UNIVERSIDAD DE MÁLAGA

E.T.S. INGENIERÍA INFORMÁTICA
UNIVERSIDAD DE MÁLAGA

Ingeniería del Software

CIENCIA DE DATOS PARA EMPRESAS TIPO AUTOESCUELA - DATA SCIENCE FOR DRIVING LEARNING COMPANIES

Realizado por
CARLOS LOBATO PADILLA

Tutorizado por
ENRIQUE ALBA TORRES
Cotutorizado por
JESÚS GABRIEL LUQUE POLO

Departamento
LENGUAJES Y CIENCIAS DE LA COMPUTACIÓN
UNIVERSIDAD DE MÁLAGA

MÁLAGA, DICIEMBRE 2020

UNIVERSIDAD DE MÁLAGA

ESCUELA TÉCNICA SUPERIOR DE INGENIERÍA INFORMÁTICA
GRADO EN GRADUADO EN INGENIERÍA DEL SOFTWARE

CIENCIA DE DATOS PARA EMPRESAS TIPO AUTOESCUELA

DATA SCIENCE FOR DRIVING LEARNING COMPANIES

Realizado por
Carlos Lobato Padilla
Tutorizado por
Enrique Alba Torres
Cotutorizado por
Gabriel Jesús Luque Polo
Departamento
Lenguajes y Ciencias de la Computación.

UNIVERSIDAD DE MÁLAGA
MÁLAGA, DICIEMBRE 2020

Fecha defensa:
El Secretario del Tribunal

Agradecimientos

A Francisca Ana Padilla Ruiz y Juan Antonio Lobato Fernández, por su incondicional apoyo y confianza plena en mi.

A Paula Lobato Padilla, por esas palabras que me han ayudado a sintetizar todo y no verlo tan complejo.

A María Sierra Triviño, por darme esa fuerza cuando estaba más bajo, ánimo y ayuda.

A Enrique Alba Torres y Gabriel Jesús Luque Polo, por aceptar, acompañarme y guiarme en esta recta final.

A AMALAJER, porque si no existiera quizás no estaría escribiendo estas palabras.

A Mariluz Padilla Ruiz, porque su recuerdo y las miradas que me dejó me dieron fuerzas para todo el camino. Espero lo estés viendo.

Resumen:

En la actualidad, las empresas, organizaciones, administraciones e invidividuos generan una cantidad de datos macroscópica. Más de 277.000 personas postean una historia al minuto. La información es uno de los activos más valiosos tanto a nivel estratégico como económico. “El conocimiento es poder´´, Francis Bacon.

Este nuevo mercado en constante avance y con tanto potencial es una de las nuevas necesidades de las empresas que quieren poder competir cara a cara con sus rivales. Productividad, efectividad, manejo de errores, predicciones y entedimiento del negocio son algunas de las ventajas que aportan el tratamiento de la información.

A partir de esta línea tecnológica de estudio y su necesidad, surge la propuesta de realizar un trabajo fin de grado sobre un proyecto de ciencia de datos. La siguiente elección fue realizarlo con una empresa y usar datos reales, con las implicaciones que conlleva.

Este proyecto pretende ayudar a Torcal Formación, la empresa seleccionada, ofreciéndole información valiosa sobre el negocio en el pasado, el presente y el futuro. Para ello se utilizarán estadísticas generales sobre la empresa, una visión basada en algoritmos inteligentes sobre el camino que sigue, ayudando a la toma de decisiones sobre los posibles pasos a dar mediante predicciones.

Además, con la información y metodologías usadas en este TFG prentedemos que la empresa Torcal Formación entienda cómo podemos ayudarles para que ofrezcan a sus clientes mejores prestaciones, servicios, personalización y comodidad.

Para conseguirlo utilizaremos procesos utilizados en el almacenamiento de datos y combinaremos técnicas y algoritmos de aprendizaje máquina, apoyados por una metodología de desarrollo definida y distintas herramientas software.

Palabras claves: Ciencia de Datos Inteligencia Artificial, Aprendizaje Máquina, Inteligencia de Negocio y Empresas

---

Abstract:

Currently, companies, organizations, administrations and individuals generate a massive amount of data, more than 277.000 people are posting one story per minute. Information is one of the most valued assets both at a strategic and economic level. "The acknowledge is power", Francis Bacon.

This new market in constant advance and with so much potential, it is one of the new needs of companies that want to be able to compete face to face with their rivals. Productivity, effectiveness, error handling, predictions, business understanding are some of the advantages provided by the information processing.

From this relatively recent technological line of study and this need, the proposal arises to perform an end-of-degree project on a data science project. The next choice was to carry out it with a company and use real data with the implications that it entails.

This project aims to help Torcal Formación, the selected compay, by offering valuable information about it´s business from the past, present and future. For this, general statistics about the business will be used, a vision based on intelligent algorithms on the path that follows to help decision-making on the possible steps to take throught by predictions.

In addition, with the information and methodologies used in this TFG, we intend that the Torcal Formación company understand how we can help them to offer thier customers better benefits, services, personalization and comfort.

To achieve this we will use processes used in data storage and we will combine machine learning techniques and algorithms, supported by a defined development methodology and different software tools.

Keywords: Data Science, Artificical Inteligence, Machine Learning, Business Inteligence and Companies

Índice de contenidos

Índice de figuras

Índice de tablas

Lista de códigos

CAPÍTULO 1

Introducción

1.1. Motivación

La ciencia de datos es una técnica que se lleva desarrollando desde hace 7 milenios cuando en Mesopotamia empezaron a usar la contabilidad para el rebaño y las cosechas. En el siglo XX empezó a tratarse por la comunidad científica sin llegar a definirlo con un nombre concreto hasta que en 2005, Roger Mougalas director de marketing de O´Reilly Media, usó el término de Big Data, refiriéndose al gran volumen de datos generados. A partir de ese momento esta ciencia ha evolucionado exponencialmente apoyándose en este término. En la actualidad, existe una gran cantidad de datos [1] en los desarrollos e investigaciones actuales, es la organización, gestión y uso de los datos por parte de algoritmos [2] lo que genera servicios y da nuevas oportunidades para aprendizaje, desarrollo de aplicaciones, soluciones, modelos predictivos, etc...

Las diferentes técnicas algorítmicas [3] (algoritmos evolutivos, de aprendizaje, técnicas paralelas) dotan de valor y sentido a los datos en crudo, recogidos por empresas y particulares, produciendo beneficios a diferentes estamentos sociales.

Figura 1.1: El círculo de la ciencia de datos

La aplicación de la ciencia de datos y la inteligencia artificial (IA) [4] es una práctica generalizada hoy en día. La aplicación de técnicas IA, estadística y conocimientos informáticos (en su mayoría software) a los sistemas de información de una empresa, se ha vuelto una tarea necesaria. Su uso conduce a mejores decisiones, predicciones y movimientos estratégicos por parte de la empresa, que en muchos casos permite marcar la diferencia a su favor contra la competencia en un mercado fluctuante, veloz y difícil.

En la actualidad una empresa que no esté beneficiándose de la ciencia de datos o las empresas que no los usen cae en la posibilidad de estancarse y quedar obsoleta. Las posibilidades en la ciencia de datos y su versatilidad son casi infinitas pudiendo mejorar cualquiera de los áreas de un negocio en el que los integrantes de este analicen que necesitan una inyección de apoyo, mejorando los resultados del negocio pasando a ser más productivo o eficiente.

Alguno de los resultados que puede proporcionar la ciencia de datos en una empresa descubriendo la información valiosa que alberga son:

1. Acelerar la innovación.
2. Obtener valor añadido en sus productos y servicios.
3. Aumentar la productividad
4. Mejorar la competitividad.
5. Acelerar el proceso de crecimiento y expansión.

En la siguiente imagen encontramos cuatro formas [5] en las que la ciencia de datos puede agregar valor a una empresa, pudiendo aprovecharlas para adelantar a la competencia:

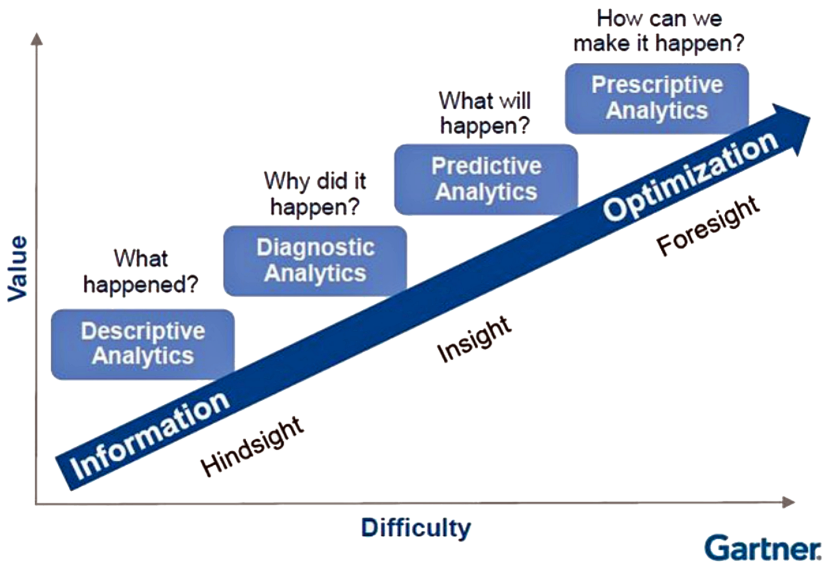

Figura 1.2: Proceso de revalorización de una empresa con ciencia de datos

Ahora describiremos en detalle cada una de las etapas en las que la ciencia de datos puede aportar valor:

Análitica descriptiva (¿Qué esta sucediendo?): El primer paso en el proceso de revalorización de la empresa, en el que se detalla una visión actual general del desempeño de tu empresa, aumentando el conocimiento sobre los clientes, trabajadores y el negocio.

- Visualizar de manera efectiva principales métricas de la empresa.
- Resumen de los datos de manera precisa y en tiempo real.

Análitica diagnóstica (¿Por qué esta sucediendo?): Después de definir el escenario de lo que sucede en la empresa, comienza el paso de formular preguntas sobre el negocio. Comprendiendo lo que ocurre alrededor de los datos, comparando y asociando resultados.

- Encuentra el motivo del problema.
- Encuentra el patrón detrás de datos y elimina el ruido.

Análitica predictiva (¿Qué es probable que suceda?): Esta etapa se basa en la creación de modelos para poder predecir movimientos futuros en base al histórico de datos recogidos.

- Uso de datos históricos de la empresa para predecir resultados específicos.
- Utilizar algoritmos para automáticamente pronosticar respuestas de clientes, resultados y tendencias.

Análitica prescriptiva (¿Qué debo hacer?): El último paso de la revalorización del negocio donde nos centramos en obtener recomendaciones para mejorar y alcanzar los objetivos que se propone la empresa, esta etapa se debe enfocar para poder beneficiar a la empresa si las predicciones eran negativas mostrar como la toma de decisiones contraria podría potenciar nuestro negocio o avisos del destino al que se puede llegar si se sigue esa trayectoria.

- Evaluación de estrategias variadas para lograr los objetivos de la empresa.
- Elegir la mejor opción y hacer recomendaciones al respecto.

La ciencia de datos abarca un gran conjunto de metodologías y cada una tiene sus beneficios y ventajas pero a groso modo todas siguen un patrón en la división de fases en el proceso de ejecución de los proyectos.

1.2. Objetivos

Este Trabajo de Fin de Grado (TFG) tiene como objetivo principal conectar al alumno a nivel laboral con el mundo real, un proyecto de ciencia de datos a una empresa real, realizando actividades con datos de usuarios a través de empresas que proporcionan la información y estén interesadas en este tipo de estudios.

Por ende, los objetivos pasan a ser la realización de unos estudios amplios y rigurosos, un análisis inteligente de esos datos mediante técnicas de análisis como el aprendizaje computacional proporcionando un conocimiento muy valioso para mejorar el servicio, la atención al cliente y procesos internos de la empresa. Todo esto se realizará con la máxima profesionalidad y guiándonos por los estándares de buenas prácticas. Para lograrlo, es necesario cumplir objetivos más concretos:

- Realizar un estudio profundizado sobre la ciencia de datos, inteligencia artificial, la empresa y sus necesidades.
- Estudiar y seleccionar las diferentes herramientas y tecnologías a utilizar durante el proceso.
- Estudiar y seleccionar metodologías de desarrollo y estructuras de los proyectos de ciencia de datos.
- Asegurarse de que la entrega de los datos cumple la protección de acuerdo con las normativas vigentes.
- Aplicación del framework OSEMN, con nuestro enfoque personalizado, usando técnicas de agrupación tradicionales K-Means/SVM, técnicas de optimización combinada para hacer selección de características, caracterización de los datos estadísticamente, programación distribuida para acelerar el cómputo y se darán resultados que serán evaluados por los Product Owners.
- Comunicación de los resultados obtenidos con una exposición clara, entendible y dinámica.

1.3. Estructura de la memoria

La memoria está estructura en dos grupos. La primera parte esta estructurada por capítulos donde se ubica al lector en el tema a tratar y las fases del desarrollo del proyecto:

1. **Introducción**: Orientar sobre la elección del tema elegido y metas a alcanzar.
2. **Fundamentos de la ciencias de datos**: Introducir la ciencia de datos tratando las bases que fundamentan este campo y dar a conocer las distintas líneas teórico-prácticas de la informática en las que se basa el trabajo fin de grado.
3. **Metodología de trabajo**: Mostrar en detalle las distintos procedimientos para desarrollar proyectos de ciencias de datos.
4. **Caso de uso: Empresa Torcal Formación**: Introducir el motivo de elegir un caso real, la empresa elegida, detallar los motivos de la elección, los beneficios que

esto aporta y una breve descripción de los datos ofrecidos.

5. **Tecnologías y herramientas usadas para este TFG**: Estudio sobre las distintas tecnologías y herramientas utilizadas en el proyecto.
6. **Implementación de un proyecto real**: Desglose detallado de las distintas fases del proyecto con las tareas, limitaciones, técnicas utilizadas.
7. **Resultados**: Análisis extenso sobre conclusiones, información relevante, predicciones, modelos, recomendaciones.
8. **Conclusiones y líneas futuras**: Análisis a nivel personal del trabajo realizado, mostrando además los caminos posibles a seguir para mejorar o ideas que se podrían desarrollar a partir de este.

La memoria contiene además una serie de índices de contenido, de figuras, de tablas, de códigos y de referencias para resultar pedagógica y fácil de reproducir en un fúturo.

Además con objeto de cumplir lo estipulado en las reglas sobre posicionamiento de cada capítulo correctamente en hojas impares, se dejarán hojas en blanco entre los finales e inicios de cada apartado que debe empezar en número impar.

CAPÍTULO 2

Fundamentos de la ciencia de datos

2.1. Ciencia de datos

Antecedentes históricos

Durante las últimas décadas ha habido un gran progreso en el campo de la tecnología y la información, con un aumento exponencial de la tecnología, las máquinas, la información... *Data* y *Analytics* son dos de las palabras más usadas en el ámbito de la tecnología y la Informática.

El término ciencia de datos empezó a tratarse hace más de 30 años, en 1962 [6], donde se relacionó la estadística con el análisis de datos, aunque no fue hasta el 1974 [7] donde se utilizó este término por primera vez.

En la década de los 90s esta área irá cogiendo cada vez más atención por parte de las empresas en su búsqueda incansable de la expansión, realizando descubrimientos sobre la minería de datos, añadiendo pasos importantes en el proceso KDD (*Knowledge Discovery in Databases*) como la preparación de los datos, selección de los datos, limpieza de los datos (*data cleaning*), etc. para asegurarse de que el conocimiento era extraído de los datos.

La nueva era de la ciencia de datos, en los inicios del siglo XXI se empezó a practicar academicamente y profesionalmente mostrando esta nueva disciplina. En el ámbito tecnológico se estaba comentando los progesos de esta nueva ciencia hasta que en 2013, IBM reportó que el 90 % de los datos a nivel mundial se habían generado en los dos últimos años.

Para que las empresas puedan beneficiarse de las ventajas de la acumulación de los datos es vital utilizar técnicas de inteligencia artifical para poder recuperar los conocimientos necesarios para que las empresas se puedan beneficiar para progresar.

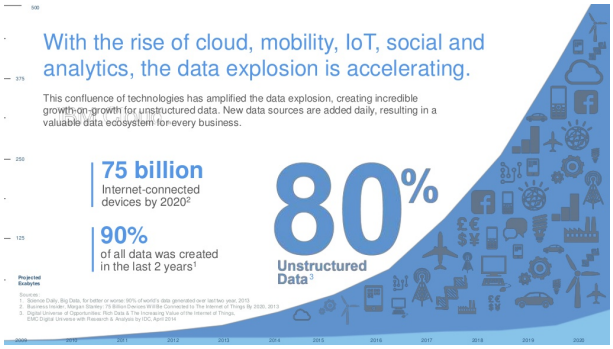

Figura 2.1: Estadística IBM sobre el almacenamiento de datos

Desde entonces la carrera por obtener y manejar la información no ha parado. La importancia de transformar grandes conjuntos de datos en información usable y encontrar patrones útiles se convirtió en una prioridad.

¿Qué es la ciencia de datos?

La ciencia de datos es la disciplina que estudia los datos. Relacionado con el concepto de ***Big Data*** (datos masivos), lo primero es la importancia de utilizar ese conocimiento, no solo de almacenarlo, al tener tantas capas de información, disponer de una serie de técnicas, métodologías y procesos de ingeniería habilitan que estas capas se dividan, se clasifiquen, se normalicen y se puedan utilizar correctamente.

A diferencia de otras técnicas de análisis de datos, ésta se puede aplicar tanto a datos estructurados como no estructurados. La ciencia es una forma más profunda y detallada de analizar datos que el análisis de datos puro, utilizando diferentes aplicaciones y herramientas, como el aprendizaje automático y algoritmos sofisticados.

Además de la explicación de los datos históricos, se utiliza análisis exploratorios y técnicas predictivas para obtener nueva información y predecir eventos futuros (Análisis causal-predictivo). También este conocimiento se aprovecha para analíticas prescriptivas, donde modelos inteligentes toman sus decisiones y aprenden mediante parámetros dinámicos de entrada.

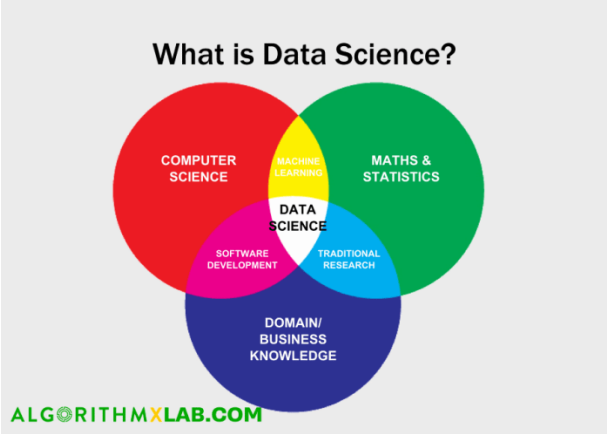

Figura 2.2: Imagen que muestra donde se sitúa la ciencia de datos, como intersección de tres dominios.

¿Por qué es necesaria?

Como la introducción histórica presenta, en el pasado los datos que acumulabamos eran reducidos y estructurados, pudiendo analizarlos manualmente o con algoritmia simple. En la actualidad la producción de datos debido a la revolución tecnológica es inmanejable, además de que los datos generados tienen estructuras semiestructuradas o totalmente desestructuradas en su mayoría, hecho que confirma la estadística presentada por IBM añadida anteriormente.

¿Para qué se utiliza la ciencia de datos?

Las utilidades detectadas con anterioridad son útiles en una infinidad de escenarios. Desde campañas de publicidad para captar la atención del usuario hasta el desarrollo de soluciones a enfermedades como el cáncer.

En la introducción de este trabajo fin de grado se centró en la revalorización del negocio ya que trabajaremos con una empresa específica pero la lista de aplicaciones [8] es enorme y a continuación mostramos una parte de las industrias que se están beneficiando de está ciencia:

- **Industria bancaria**
- **Industria de viajes**
- **Comercio minorista**
- **Redes sociales**
- **Vehículos autónomos**
- **Servicios de salud**

2.2. Matemáticas

Como vimos previamente en la Figura 2.2, uno de los dominios que dan lugar a este campo es las Matemáticas, y este apartado veremos varios elementos de los que se nutre la ciencia de datos.

2.2.1. Álgebra lineal

El álgebra lineal es la rama de las matemáticas relacionada con las estructuras matemáticas bajo operaciones de suma y multiplicación escalar incluyendo la teoría de sistemas de ecuaciones lineales, matrices, determinantes, espacios vectoriales y transformaciones lineales.

Es uno de los requísitos para trabajar en el ámbito de la ciencia de datos, enfocada a el entendimiento de cómo funcionan los algoritmos utilizados en el aprendizaje máquina (*machine learning*) y profundo (*deep learning*). Determinaremos varias aplicaciones en la ciencia de datos:

Funciones de pérdida:

Una función de pérdida es la aplicación del vector normalizado en álgebra lineal, pudiendo ser simplemente su magnitud. Estas funciones se usan para calcular las diferencias entre las predicciones y las sálidas esperadas. A modo de ejemplo, aquí se muestran dos de ellas:

- Distancia Manhattan:

$$\begin{gathered}\mathbf{d(p,\ q)} = ||p - q|| = \sum_{i=1}^{n} |p_i - q_i| \\ \text{donde } \mathbf{(p,\ q)} \text{ son vectores} \\ p = (p_1, p_2, ..., p_n) \text{ y } q = (q_1, q_2, ..., q_n)\end{gathered} \tag{2.1}$$

- Distancia Euclídea:

$$\begin{gathered}\mathbf{d(p,\ q)} = \sqrt{\sum_{i=1}^{n} (q_i - p_i)^2} \\ \text{donde } \mathbf{(p,\ q)} \text{ son vectores} \\ p = (p_1, p_2, ..., p_n) \text{ y } q = (q_1, q_2, ..., q_n)\end{gathered} \tag{2.2}$$

Regularización

Técnica utilizada para prevenir que los modelos se sobreadapten. Un modelo se sobreadapta cuando se ajusta al entrenamiento de los datos demasiado correctamente y luego no es capaz de trabajar adecuadamente con datos no mostrados previamente.

La regularización penaliza los modelos demasiado complejos añadiendo la norma del peso del vector a la función de coste. Al buscar minimizar esta función de coste se quedrá minimizar esta norma.

Principalmente se pueden considerar dos tipos de regularización:

- Regressión *Lasso*
- Regresión *Ridge*

Matriz de covarianza

El análisis multivariable es un paso importante en la exploración de los datos, para estudiar la relación entre un par de variables. La covarianza y la correlación son medidas que estudian las relaciones entre dos variables continuas. En concreto, la covarianza nos indicará la dirección de la relación lineal entre dos variables:

- Covarianza positiva: El crecimiento o decrecimiento estará acompañado por la otra variable.
- Covarianza negativa: El crecimiento o decrecimiento será contrario al de la otra variable.

TABLE 3.4. *Estimators of β_j in the case of orthonormal columns of* $\mathbf{X}$. *M and λ are constants chosen by the corresponding techniques; sign denotes the sign of its argument (± 1), and x_+ denotes "positive part" of x. Below the table, estimators are shown by broken red lines. The 45° line in gray shows the unrestricted estimate for reference.*

Estimator	Formula
Best subset (size M)	$\hat{\beta}_j \cdot I(\lvert\hat{\beta}_j\rvert \geq \lvert\hat{\beta}_{(M)}\rvert)$
Ridge	$\hat{\beta}_j/(1+\lambda)$
Lasso	$\text{sign}(\hat{\beta}_j)(\lvert\hat{\beta}_j\rvert - \lambda)_+$

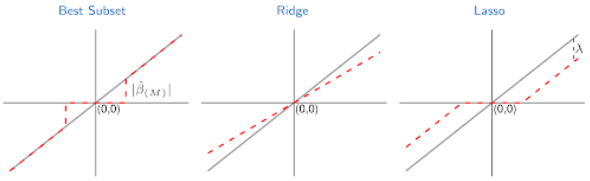

Figura 2.3: Gráficas de las regresiones Lasso y Ridge

Usando el concepto de transformación y multiplicación matricial podremos obtener resultados:

$$cov = X^t X \tag{2.3}$$

siendo **X** la matriz estandarizada de datos que contiene todos los valores numéricos.

Máquinas de vectores de soporte

Uno de los algoritmos de clasificación más usados es la aplicación del concepto de espacios vectoriales en álgebra lineal.

Las máquinas de vectores de soporte, o más conocido como *Support Vector Machine (SVM)*, es un clasificador discriminativo que funciona encontrando una superficie de decisión [9].

En estos algoritmos tendremos espacios n-dimensionales donde n es el número de valores que son parte del espacio de búsqueda. Los hiperplanos son la clave para encontrar la diferenciacion entre las dos clases. Usaremos las transformaciones de Kernel.

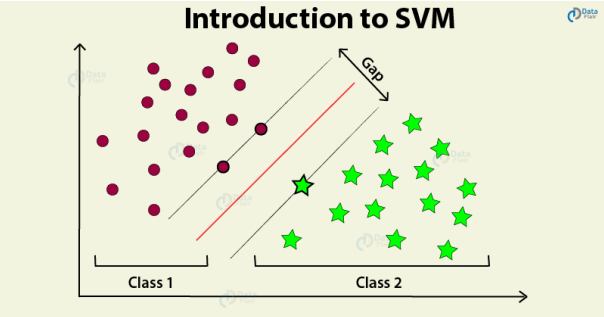

Figura 2.4: Esquema explicativo de máquinas de vectores de soporte.

2.2.2. Cálculo

Cálculo es la rama de las matemáticas que estudia el ratio de cambio de cantidades (normalmente representadas con curvas) y el tamaño, área y volumen de objetos.El cálculo se suele dividir en dos grandes secciones:

- Diferenciación: Divide en pequeñas partes para encontrar como cambia.

$$\frac{df}{dx}x = 1 \tag{2.4}$$

- Integración: Une las pequeñas partes para encontrar cuanto realmente hay.

$$\int 1dx = x + C \tag{2.5}$$

Aquí nos centraremos al igual que se hizo en el álgebra lineal a la formulación de funciones usadas para entrenar a los algoritmos utilizados en el aprendizaje máquina (*machine learning*) y profundo (*deep learning*) en la búsqueda de alcanzar sus objetivos. Determinaremos varias aplicaciones en la ciencia de datos:

Gradiente descendiente

En nuestro modelo el objetivo es reducir el coste en los datos de entrada. La función de coste se usa como monitorizadora del error en las predicciones de un módelo de aprendizaje máquina.

La minimización de este coste tiene como objetivo encontrar el menor valor del error posible e incrementa la corrección del modelo. Esto último se consigue entrenando el modelo mediante la modificación de los parámetros de nuestro modelo.

Lo tradicional para obtener esta información es definir una función de error o **función de coste**, donde se irán seleccionando pares de valores (real y modelado) para devolver el error producido por nuestro modelo y así comprobar cuanto se adapta nuestro modelo a los datos.

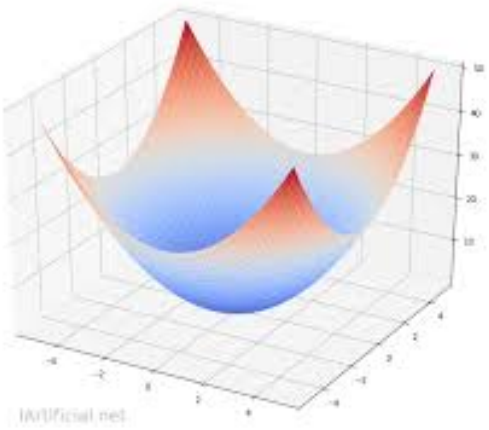

Figura 2.5: Ejecución del gradiente descendiente

Regresión de los mínimos cuadrados

El método utilizado para esta función de coste es el popular método de los mínimos

cuadrados.

$$F(x_i) = \frac{1}{N}\sum_{i=1}^{N}(y_i - f_i)^2 \quad \text{siendo } f(x_i) = mx_i + b \tag{2.6}$$

Nuestro gradiente funcionará como un subdivisor claro, por lo que necesitaremos calcular este gradiente para encontrar el error que estará divido por dos parámetros, usaremos derivadas parciales para cada uno:

$$\frac{\partial}{\partial m} = \frac{2}{N}\sum_{i=1}^{N} -x_i(y_i - (mx_i + b)$$
$$\frac{\partial}{\partial b} = \frac{2}{N}\sum_{i=1}^{N} -x_i(y_i - (mx_i + b)$$

Para finalizar dejaremos que nuestro algoritmo de descenso del gradiente se ejecute y encuentre la opción con menor error, mediante pasos iterativos guiado por los errores encontrados.

2.2.3. Estadística

Estadística se define como la rama de las matemáticas formada por un conjunto de métodos científicos ligados a la toma, organización, recopilación, presentación y análisis de una serie de datos, tanto para la deducción de conclusiones como para tomar decisiones razonables de acuerdo a esos análisis permitiendo comprendender un fenómeno en particular.

Es decir, es el dominio a través del cual se recolecta, analiza, describe y estudia una serie de datos a fin de establecer comparaciones o varaibilidades que permitan comprender un fenómeno en particular. Podemos distinguir dos categorías:

- **Estadística descriptiva:** Utiliza los datos para proveer descripciones sobre población, cálculos numéricos, grafos o tablas de manera informativa.
- **Estadística inferencial:** Determina propiedades, crea inferencias y predicciones sobre la población basándose en los datos recopilados por esta población con base en una muestra de ella.

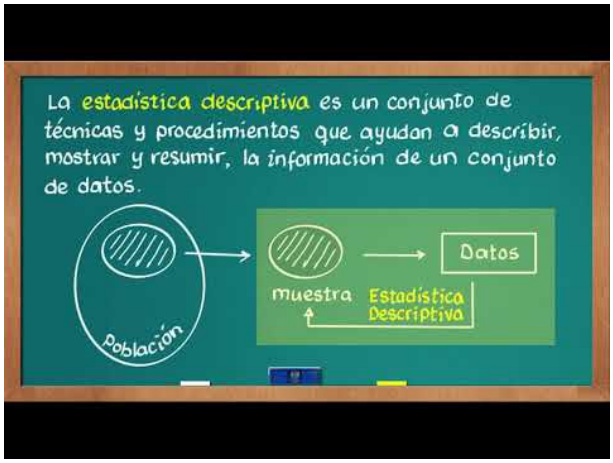

Figura 2.6: Esquema de la estadística descriptiva

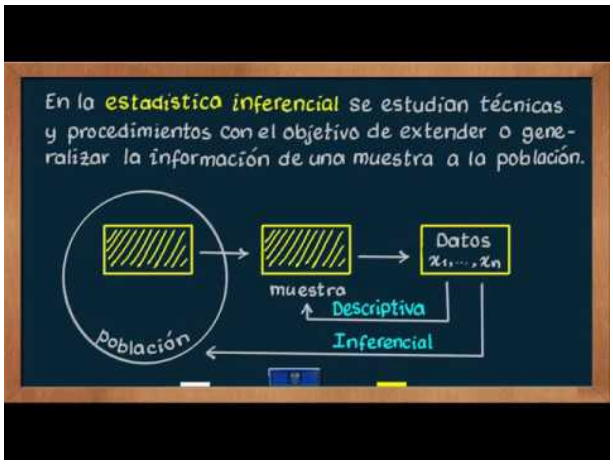

Figura 2.7: Esquema de la estadística inferencial

Paradigma Bayesiano

- La probabilidad se describe en grados de creencia, no frecuencias límite, pudiendo hacer afirmaciones probabilística de cualquier tipo y no solamente de datos sujetos a variabilidad aleatoria.

- Está permitido realizar afirmaciones probabilísticas de parámetros.

- Está permitido inferenciar parámetros por medio de distribuciones de probabilidad, pudiéndose extraer estas en forma de estimaciones de intervalos y puntuales de dicha distribución.

Regla de Bayes en modelos y datos

Usada para determinar la probabilidad de un modelo dado un conjunto de datos. El modelo determina la probabilidad de los datos condicionado a valores particulares y la estructura del modelo. La regla de Bayes inferirá la probabilidad del modelo usando los

datos, desde la probabilidad de los datos, dado el modelo.

$$P[A_n/B] = \frac{P[B/A_n] \cdot P[A_n]}{\sum_{i=1}^{k} P[B/A_I] \cdot P[A_i]} \tag{2.7}$$

donde $P[A_n]$ son las probabilidades a priori

Regresión lineal

La regresión lineal es un método para predecir variables objetivos ajustando la mejor relación lineal entre las variables dependientes e independientes. El mejor ajuste se realiza asegurándonos de que la suma de todas las distancias entre las esperadas y las observaciones actuales es lo mínimo posible. Cuando ninguna otra posición pueda dar mejor mínimo el ajuste será el mejor posible.

- Regresión lineal simple: Este tipo de regresión usa una variable independiente para predecir una variable dependiente ajustando la mejor relación lineal.
- Regresión lineal múltiple: Este tipo de regresión usa más de una variable independiente para predecir una variable dependiente ajustando la mejor relación lineal.

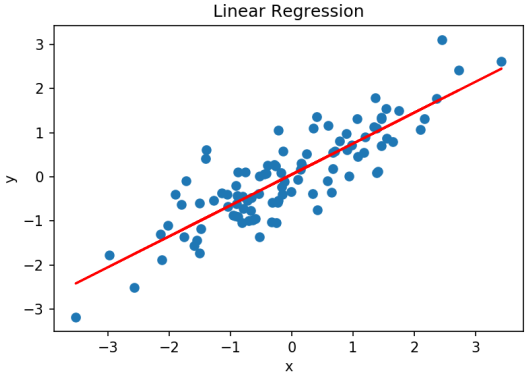

Figura 2.8: Ejemplo de regresión lineal

2.3. Inteligencia Artificial

La inteligencia artficial, rama de la informática, con fuertes raíces en otras áreas como la lógica y las ciencias cognitiva, se define como la ciencia e ingenería de crear máquinas inteligentes, especialmente enfocado a aplicaciones informáticas mediante los cuales validar el trabajo realizado.

Nace en Dartmouth (Estados Unidos) en 1956 donde participaron los investigadores principales del área, J. McCarthy, M. Minsky, N. Rochester y C. E. Shannon, primeros científicos en utilizar la palabra. Este documento [10] definía el problema de la inteligencia artificial como aquel de construir una máquina que tuviera un comportamiento igual que el humano, llamándolo inteligente. A raíz de este debate se generan definiciones no basadas en el comportamiento humano:

1. Actuar como las personas: Modelo a seguir para la evaluación de los programas corresponde al comportamiento humano, usado en el famoso Test de Turing (1950)

2. Razonar como las personas: El razonamiento por encima del resultado de este.
3. Razonar parcialmente: El razonamiento como base del modelo pero con la premisa de un forma racional de razonar preconcebida.
4. Actuar racionalmente: Los resultados como foco, pero evaluándolos de forma objetiva.

Campos en la Inteligencia Artificial

Encontramos una serie de técnicas, métodos desarrollados y resultados atribuibles a esta rama de la informática:

- **Resolución de problemas y búsqueda:** El objetivo principal de la inteligencia artificial es resolver problemas de índoles diversas, se necesitará formalizar esos problemas para resolverlos.
- **Representación del conocimiento y sistemas basados en el conocimiento:** Comúnmente en inteligencia articial se necesita incorporar conocimiento del dominio de aplicación para la resolución de problemas.
- **Aprendizaje automático:** El rendimiento de cualquier máquina se incrementa si la máquina aprende de la actividad realizada y sus errores.
- **Inteligencia artificial distribuida:** Versiones paralelas de métodos ya existentes o problemas con los agentes autónomos.

Figura 2.9: Áreas de la inteligencia artificial

2.3.1. Machine Learning

Machine learning [11] se define como el campo de estudio que otorga a las máquinas la habilidad de aprender sin que se les programe para ello. De este modo, este campo se centra en la búsqueda de patrones para hacer previsiones y en el desarrollo de programas que accedan a la información y la usen para aprender por si mismos.

Este proceso comienza con las observaciones de los datos, tales como ejemplos, experiencia directa o instrucciones, como parte clave de la búsqueda de patrones en los datos y realizar mejores decisiones en un futuro basado en los elementos introducidos al programa. El principal objetivo como ya comentamos es permitir que los ordenadores aprendan automáticamente sin intervención humana o asistencia ajustando las acciones acordes.

- **Aprendizaje supervisado**

Los modelos de aprendizaje supervisado [11] son aquellos en los que se aprenden funciones, relaciones que asocian entradas con salidas, ajustándose al conjunto de elementos introducidos de los que conocemos la relación entre la entrad y la salida deseada.

Este hecho llega a proporcionar una de las clasificaciones más habituales en el tipo de algoritmos que se desarrollan, gracias a esto, dependiendo del tipo de salida que obtengamos encontramos dos subdivisiones importantes.

Modelos de clasificación: Esta subcategoría se caracteriza porque la salida es un valor categórico.

Modelos de regresión: Esta subcategoría se caracteriza porque la salida es un valor de un espacio continuo.

- **Aprendizaje no supervisado**

Los modelos de aprendizaje no supervisado [11] son aquellos en los que nuestro objetivo no es ajustar pares de entrada y salida, sino aumentar el conocimiento estructural de los datos disponibles y futuros que prosigan el mismo patron.

La utilización de distintas técnicas normalmente siguen un patrón general en el que dada una agrupación de los datos según sus semejanzas utilizando *clustering*, simplificamos la estructura de los datos manteniendo las características fundamentales como en procesos de reducción de la dimensionalidad o extrayebnndo las estructura interna con la que se distribuyen los datos en su espacio original, aprendizaje topológico.

- **Aprendizaje reforzado**

Los modelos de aprendizaje reforzado son aquellos en los que nuestro objetivo es obtener un algoritmo que aprenda de su propia experiencia. El algoritmo deberá poder tomar la mejor decisión en diferentes situaciones basándose en el proceso de prueba y error en el que la mejor decisión es premiada. Utilizado en reconocmiento facial, diagnósticos médicos o clasificación de secuencias de ADN.

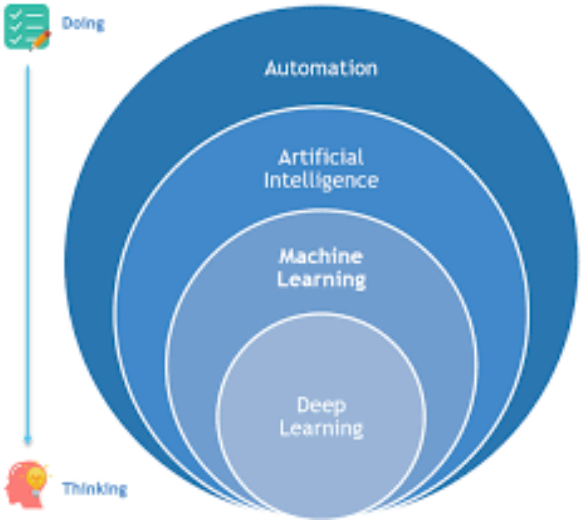

Figura 2.10: Subdivisión de las partes de la inteligencia artificial

2.3.2. Deep Learning

Está relaciando con redes neuronales artificiales que están compuestas por muchas capas. Particularmente abarca la alimentación utilizando redes neuronales de un sistema informático compuesto por una gran cantidad de datos pudiendo ser usados para tomar decisiones sobre otros datos.

El deep learning o aprendizaje profundo es una técnica que facilita la implementación del *machine learning*, produciendo una estrecha relación entre las tecnologías. En los sistemas de deep learning, al igual que en otros sitemas supervisados, es preciso indicar que deben hacer. Este tipo de sistemas, gracias a su múltiples capas ocultas, está pensado para trabajar con grandes cantidades de datos.

CAPÍTULO 3

Metodología de trabajo

3.1. Método científico

El método científico es el método de investigación para el concomiento de la realidad observable, que consiste en formularse interrogantes sobre esta realidad, con base en teoría existente, tratando de hallar soluciones a los problemas planteados. El método científica se basa en la recopilación de datos, su ordenamiento y su posterior análisis.

Todo proyecto de ciencia de datos si nos abstraemos de conceptos específicos puede analizarse como un proyecto científico y que mejor base de trabajo que una metodología tan extendida y ratificada durante tantos años.

El método más utilizado por la comunidad científica es el hipotético-inductivo, comenzando de la observación, elabora un modelo interpretativo de la información recabada de los hechos observados y luego itera corrigiendo el modelo inicial a partir de nuevas observaciones.

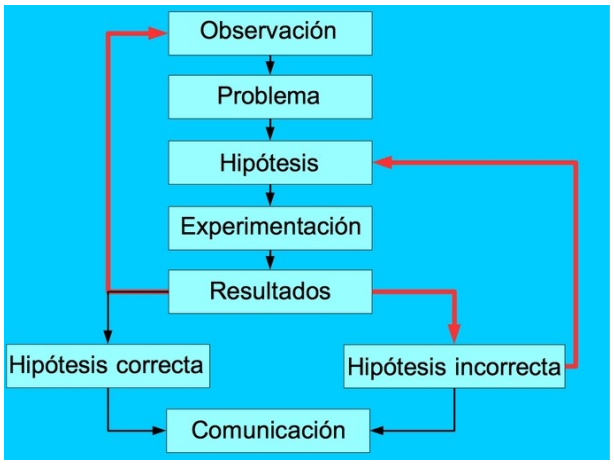

Figura 3.1: Flujo del método científico

Puede además experimentar, provocando situaciones nuevas relacionadas con los procesos naturales analizados, permitiendo obtener nuevas informaciones que transcienden la mera observación. Siendo la observación y el registro de fenómenos naturales la piedra angular del método.

La verdad científica se basa en una fusión entre realidades naturales y teorías científicas que explican su razón de ser. El científico formula una generalización, una hipótesis científica, basada en sus observaciones, por esta razón, le permite hacer predicciones. Estas predicciones serán comprobadas mediante experimentos para determinar si el resultado esperado es obtenido. Si las predicciones coinciden con el resultado, la hipótesis será ratificada. Pasando su teoría a ser una ley científica.

Los pasos del método científico se dividen en:

1. **Observación:** La base del método científico y la fuente última de todos los descubrimientos debe ser cuidadosa, precisa, con experimentos que devuelvan resultados repetitivos, testigos adecuados y lo más cuantitativo posible. Además, deben constar de un registro que serán los datos del experimento.

2. **Hipótesis:** Guía para lo que se investiga, explicaciones tentativas del fenómeno investigado y se formulan como proposiciones acerca de las relaciones entre dos o más varaibles. Una hipótesis es una suposición activa.

3. **Experimentación:** La prueba científica de una hipótesis, debiéndose diseñar un conjunto de experimentos para probar la hipótesis propuesta. Mientras se realizan los experimentos se debera registrar la información, en los experimentos la muestra utilizada debe ser representativa. Obteniendo la cantidad de información suficiente y confiable. Una vez registrados los datos se deben organizar y analizar.

4. **Conclusiones y Teorías:** Los datos obtenidos de un experimento se analizan con la finalidad de corroborar o refutar la hipótesis original. Una hipótesis apoyada en muchas observaciones y experimentos distintos se transforma en teoría, principio general cienítifico aceptado ofrecida para explicar sucesos. Una teoría es un poder conceptual que explica las observaciones existentes y predice resultados de nuevas observaciones.

Las siguientes metodologías que se desarrollan, son las metodologías en las que hemos plasmado este método científico utilizando una serie de fases de cada una que conmutan en la realización del método científico en su totalidad.

3.2. Metodología de proyectos de ciencias de datos

Un proyecto de ciencias de datos [12] es un procedimiento de ingeniería (comienzo, pasos y final). Repleto de decisiones informadas retroalimentandose durante los distintos pasos y basándose en un criterio predefinido.

El objetivo de un proyecto de ciencias de datos es optimizar el uso de los recursos a la par que se maximizan los beneficios, la rentabilidad del negocio es necesaria pero la obtención de hipótesis e ideas reales es obligatoio.

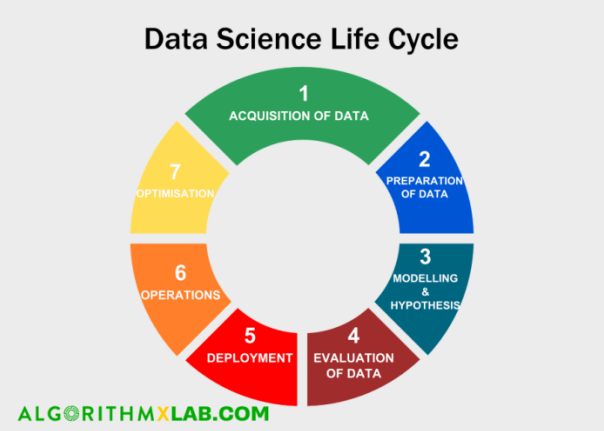

Figura 3.2: Ciclo de vida de un proyecto de ciencia de datos

Hay distintas metodologías, la más genérica y extendida está formada por siete pasos en las que se dividen el ciclo de vida de los datos:

1. **Adquisición de los datos:** Recolección de datos desde fuentes internas y externas. Incluyendo técnicas de *web scraping* o llamadas a APIS en redes sociales.

2. **Preparación de los datos**: Normalmente referido como *data wrangling*, esta fase envuelve la limpieza de los datos y transformaciones en formatos útiles para funciones en ciencia de datos. Similar a los tradicionales pasos ETL en almacenamiento de datos, pero añade más análiss y extracciones en formatos idóneos.

3. **Hipótesis y modelado:** Práctica muy común en data mining, pero en proyectos de ciencia de datos no hay límites, teniendo como objetivo aplicar técnicas de machine learning a todos los datos. La selección del modelo implica identificar conjuntos de entrenamiento para entrenar candidates a modelos machine leraning y probar esos conjuntos para el siguiente paso.

4. **Evaluación e interpretación:** Comparar el rendimiento de un modelo y seleccionar los mejores, calculando la exactitud y preveniendo la sobreadaptación

   Estos dos últimos pasos se iteraran tantas veces como sea necesario hasta tener un conocimiento claro y resultados sobre los modelos iniciales y las hipótesis hayan sido evaluadas.

5. **Despliegue:** El proyecto se ejecutará en un entorno de pre-producción antes de sacarlo a producción o permitirá cambios después del despliegue, basándonos en el modelo de despliegue continuo.

6. **Operaciones:** Esta fase irá seguida de un modelo DevOps congeniando con el modelo de despliegue. Es recomendable que el despliegue incluya tests de rendimiento para controlar que el rendimiento tenga una estabilidad.

   Los pasos cinco y seis también tendrán una iteración al igual que en el software ágil.

7. **Optimización:** El último paso del ciclo de vida de los datos. Paso al que se recurrirá si aparecen fallos de rendimiento o se necesita añadir nuevos datos y reentrenar el modelo.

Utilizaremos esta metodología como referencia a alto nivel, debido a que es más completa y tiene un enfoque más científico que las otras comentadas a continuación (OSEMN y CRISP-DM). Las dos metodologías siguientes están enfocadas a metodologías más técnicas a nivel de datos o de negocio.

Los dos primeros pasos estarán relacionados con la observación que demanda el método científico, el tercero con hipótesis, cuarto con experimentación donde probamos hasta encontrar los mejores modelos y los dos siguientes con conclusiones y teorías, ya que se afirma que los resultados son los mejores posibles y se desarrollan.

3.3. Framework OSEMN

Una buena extracción, organización y seguir una serie de procesos estandarizados como OSEMN [13] incrementa la probabilidad de realizar alisis acertados y facilita volver a un paso específico dentro de un flujo definido del procesamiento de datos.

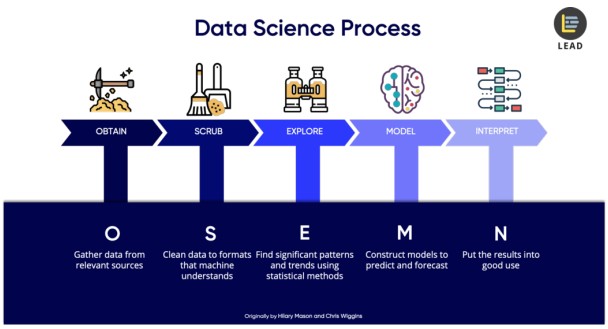

Figura 3.3: Pasos del proceso OSEMN

El proceso OSEMN es un modelo de organización, tareas para la investigación en el campo de la ciencia de datos estandarizado y mundialmente aceptado. Busca solventar el problmema con la ciencia de datos a gran escala.

Este framework ofrece una secuencia clara de actividades (Obtain, Scrub, Explore, Model y iNterpret) en las que las metas se van archivando en cada paso. Se utiliza como base en la mayoría de modelos alternativos para gestionar los proyectos de ciencias de datos.

Los pilares de nuestra metodología de trabajo a bajo nivel están basados en esta técnica tan extendida en la rama de la ciencia de datos, con un paso extra añadido al final de esta sección, paso debatido en la comunidad de científicos de datos por necesitar integrarlo. El último paso de interpretación sube al alto nivel por su importancia en la clausura del proyecto.

Obtain data - Recolección

En este paso, debemos extraer y almacenar los datos dependiendo del tipo de proyecto se realizarán unas técnicas u otras.

- Obtener los datos por parte de una empresa u organización.
- Descargar los datos desde otra localización (webs o servidores)
- Consultas desde una base de datos o API (MYSQL o Twitter)

- Extraer datos desde algún tipo de arcihvo (archivo HTML, hojas de cáclulo, CSV o JSON)
- Generar los datos por nosotros mismos (lectura de sensores, realizando encuestas, registros digitales)

Scrub data - Limpiar y normalizar

Generalmente entre los datos obtenidos se encuentran valores incompletos o vacíos, inconsistencias, errores, carácteres extraños, datos que no son interesantes, útiles o necesarios en el estudio en cuestión.

En estos casos deberemos limpiar los datos para poder devolver resultados interesantes. Lás operaciones más comunes son:

- Filtrar filas.
- Extraer columnas específicas.
- Reemplazar valores.
- Extraer palabras.
- Manejar valores vacíos.
- Convertir los datos a otros formatos.

Usualmente, la mayor parte del esfuerzo en un proyecto de ciencia de datos está enfocado en estos dos primeros pasos. En [14] se afirma que el 80 % del trabajo en cualquier proyecto de datos es la limpieza del mismo.

Explore data (EDA) - Exploración

Una vez tenemos los datos y además están limpios, estamos preparado para poder buscar conclusiones e información relevante sobre estos, es la fase donde deben aparecer las preguntas que un científico de datos se haría. Los pasos a seguir serían:

- Mirar minuciosamente los datos.
- Creación de estadísticas derivadas de los datos.
- Producir visualizaciones significativas.

Primeramente, se debe inspeccionar los datos y sus propiedades, para enfocar el manejo específico que deben tener.

Siguiente paso será crear estadísitcas descriptivas para extraer variables significativas, cualidades y funcionalidades valorables.

Finalmente, la creación de visualizaciones gracias a las cuales identificar patrones y tendencias de los datos.

Model data - Modelado

Este es el paso donde la ciencia de datos muestra todo su potencial, donde nos centraremos. La importancia de haber realizado los pasos anteriores correctamente, al igual que en la Exploración, es crucial para la construcción de modelos útiles.

Primeramente, debemos reducir la dimensión de nuestro conjunto de datos. No todos los valores serán esenciales para predecir un modelo. Debemos seleccionar los valores más relevantes.

Las técnicas utilizadas para la creación de nuestro modelo:

- Regresión.
- Clasificación.
- Predicción.
- Clustering.

Interpret the data - Interpretación

El úlltimo paso y más importante de un proyecto de ciencia de datos, interpretar los modelos y los datos. El poder de predicción de un modelo recae en su habilidad para generalizar. Nuestra explicación de este modelo depende en su habilidad para generalizar datos futuros.

- Redactar las conclusiones de nuestros datos.
- Evaluar el sentido de los resultados obtenidos.
- Comunicar los resultados.

Esta interpretación de los datos debe estar enfocada para al presentación para un público no técnico, una práctica que ayuda en este sentido es usar las preguntas de ciencias de datos generadas con anterioridad, mezclado con la información que dan la posibilidad de producir acciones claras en el futuro.

Vemos aquí como la ciencia de datos ofrece analisis predictivos y prescriptivos. Es esencial presentar los hallazgos de forma que la empresa pueda utilizarlos en su beneficio. En este apartado, las habilidades técnicas no son suficientes. Habilidades comunicativas, narrativas, creativas y sociales.

Analizando las distintas metodologías y los proyectos de ciencia de datos se hace necesario una etapa atemporal que se ejecuta durante todo el proceso para ir optimizando y readaptando el proyecto a las necesidades que surgen druante la ejecución de este.

Iterate - Iteración

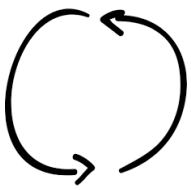

Figura 3.4: Símbolo de la iteración

El sentido de un proyecto de ciencias de datos es probar su efectividad para justificar el trabajo realizado, conseguir resultados iniciales e ir optimizándolos es una manera de tener resultados desde casi su inicio además de asegurarnos el refinamiento de toda la información obtenida. Para nuestro proyecto añadiremos este paso que es aplicable a las distintas metodologías usadas.

La relación con el método científico en esta metodología también es clara las dos primeras fases se relacionan con la observación, en la tercera empezaremos a hacernos preguntas por lo que encontrariamos la fase de hipótesis, la cuarta fase se relaciona con la experimentación ya que confirmaremos hipótesis y crearemos los modelos, la última parte se relaciona con conclusiones y teorías donde mostramos los resultados con pruebas de su efectividad.

3.4. Proceso de análisis de datos

En un proyecto de ciencias de datos hay muchas claves que tener en cuenta y que pueden perderse por el transcurso de este. Otra manera de asegurarnos de que nuestro proyecto sigue un curso óptimo es aplicar procesos aceptados a nivel mundial por el mundo empresarial y la comunidad científica.

El proceso elegido para continuar completando nuestra metodología final ha sido CRISP-DM (Cross-Industry Standard Process for Data Mining) [15] un modelo de proceso estandarizado abierto que describe enfoques generalizados usados por expertos en minería de datos, cómo su nombre indica es un proceso enfocado en la minería de datos o el análisis de datos por lo que no sería suficiente utilizar solamente estos pasos en un proyecto de ciencia de datos.

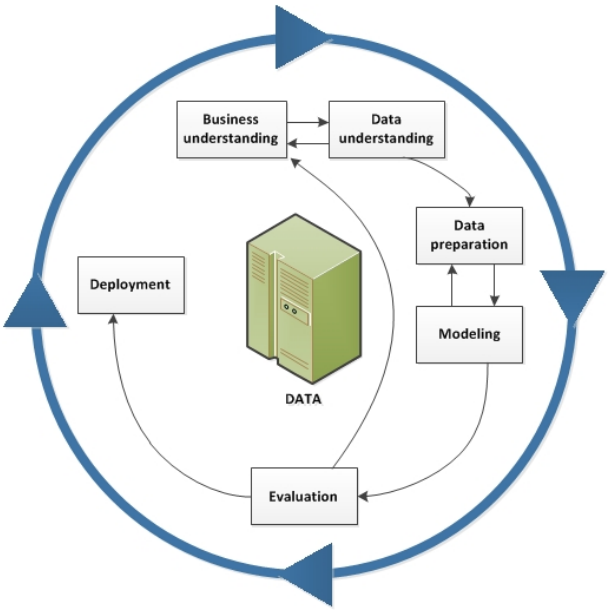

Figura 3.5: Pasos CRISP-DM creado IBM

Los pasos que más interesan al científico de datos serían los dos primeros que tienen un enfoque muy útil de cara a entender el dominio del negocio y los datos proporcionados, debido a que cuando comenzamos un proyecto de ciencia de datos y más en colaboración con una organización tenemos poco conocimiento sobre sus preocupaciones, objetivos, información interna al igual que de sus datos.

El proceso se divide en las siguientes 6 fases, nos centramos en las dos primeras porque las restantes son similares a las de las metodologías explicados en los apartados previos, se explicaran los dos iniciales ya que serán los que se apliquen en nuestro método:

1. Entendimiento del negocio: Fase que se encarga de explorar las necesidades del negocio de cara a la ejecución de la mineria de datos, relacionándose con las máximas personas clave y documentando los resultados. Finalmente se discutirá como producir un plan de proyecto usando la información rescatada.
2. Entendimiento de los datos: Fase que se enfoca en la revisión y análisis minucioso de los datos obtenidos, creando tablas y gráficas para poder determinar la calidad de los datos.
3. Preparación de los datos
4. Modelado
5. Evaluación
6. Despliegue

Siguiendo el hilo de las metodologías anteriores las tres primeras fases se relacionan con observacion, la cuarta fase se relaciona con la fase de hipótesis, experimentación se relaciona con la fase de evaluacion y despliegue al llegar a esta fase cuando todos los pasos anteriores se han ratificado estaría enlazada con conclusiones y teorías.

3.5. Estructura en entornos de proyectos de ciencias de datos

En la gestión de proyectos unificar configuraciones, estructuras y elecciones simplifica y agiliza mucho el trabajo. Los proyectos de ciencia de datos tienen unas configuraciones generales que en la sección de tencología y herramientas detallaremos aún más.

- Uso de herramiento de gestión de versiones.
- Construcción de una estructura de carpetas del proyecto estandarizada.
- Uso de entornos virtuales para su individualización de otros proyectos.
- Uso de blocs de notas donde ir recogiendo toda la información.
- Uso de *scripts* para agilizar el trabajo.
- Herramientas de *teamworking*, chat y videoconferencias.

3.6. Metodología de desarrollo

Para conseguir una organización, planificación, ejecución y control del desarrollo del proyecto efectivo es necesario definir una metodología de desarrollo desde el comienzo del proyecto. Esta metodología guiará el progreso del proyecto. Los distintos modelos de desarrollo se dividen en varios grandes bloques:

- Modelos tradicionales: Formados por un conjunto de fases o actividades en las que no tienen en cuenta la naturaleza evolutiva del software.
  - Clásico, lineal o en cascada.
  - Estructurado.
  - Basado en prototipos.
  - Desarrollo rápido de aplicaciones (RAD).
- Modelos evolutivos: Modelos que se adaptan a la evolución que sufren los requisitos del sistema en función del tiempo
  - En espiral.
  - Evolutivo.
  - Incremental.
  - Modelo de desarrollo concurrente.
- Modelos orientados a la reutilización: Enfoque de desarrollo que trata de maximizar la reutilización de software existente.
  - Basado en componentes
  - Proceso unificado
- Modelos para sistemas orientados a objetos: Modelos con un alto grado de iteratividad y solapamiento entre fases.
  - De agrupamiento
  - Fuente
  - Basado en componentes
  - Proceso Unificado
- Proceso ágiles: Nuevo enfoque en el desarrollo de software equipos pequeños, desarrollo incremental e iterativo, programación en cajas de tiempo, flexibilidad en la adopción de cambios y nuevos requisitos, colaboración e interacción constante con el cliente.
  - Programación externa (XP)
  - Desarrollo de software adaptativo
  - SCRUM
- Modelos para sistemas web: Desarrollo focalizado en sisemas y aplicaciones basados en Web.
  - UML-based Web Engineering

La opción más adecuada para nosotros ha sido SCRUM [16], además de la adecuación de la teoría de la metodología con nuestro proyecto tenemos una experiencia dilatada usándola.

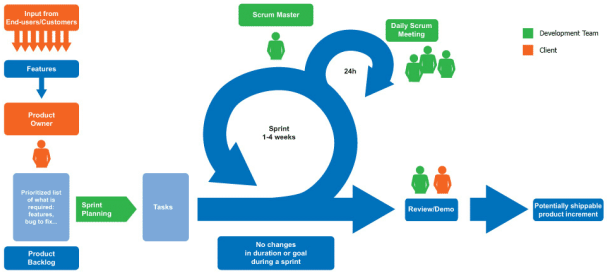

Figura 3.6: Diagrama de como aplicar SCRUM

Revisando las opciones que existen y las características específicas de nuestro proyecto la elección encaja a la perfección, en el cuál tiene una tipificación diferente al ser un proyecto de ciencia de datos, tener un equipo reducido, necesitar una interacción constante con el cliente para el entendimiento del negocio y de los datos, la importancia de iterar en los pasos del proyecto debido a las bases de los proyectos de ciencias de datos y la necesidad de optimizar los resultados, además de la debida modificación de modelos, conjunto de datos para entrenar, nuevas posibilidades dependiendo de los resultados obtenidos.

3.7. Metodología para la productividad

Las metodologías de productividad son sistemas formados por un conjunto de reglas o pasos que funcionan secuencialmente y unas serie de herramientas que acompañan y apoyan el sistema.

El método elegido ha sido Getting Things Done (GTD) [17], cimentado en cinco pasos claros y fundamentales:

- Capturar: Recopilar lo que atrae nuestra atención.
- Clarificar: Procesar el sentido de lo recolectado (accionable, innecesario, referencias, etc...)
- Organizar: Distribuir en el lugar idóneo.
- Evaluar: Revisar frecuentemente si es necesario.
- Ejecutar: Simplemente hazlo.

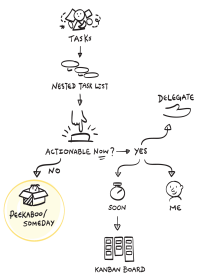

Figura 3.7: Ejecución del método GTD

La herramienta principal utilizada para aplicar esta técnica de productividad es Trello, definida en el capítulo de Tecnologías y Herramientas.

3.8. Conclusión

Después de las explicaciones de cada una de las metodologías a partir de las cuales se formaliza una estructura final de desarrollo que aplicaremos en nuestro método de trabajo.

- **Entedimiento del negocio:** Fase del desarrollo relacionada con la primera etapa del proceso de análisis de datos CRISP-DM, donde nuestro conocimiento sobre el negocio, problemáticas, necesidades y objetivos aumenta considerablemente, además de la preparación de un plan de trabajo.
- **Adquisición de los datos:** Fase del desarrollo relacionada con la primera etapa de la metodología de proyectos de ciencias de datos y con la etapa de recolección de los datos del framework OSEMN, donde obtenemos los conjuntos de datos en el entorno de trabajo tras comprobaciones sobre estos.
- **Entendimiento de los datos:** Fase del desarrollo relacionada con la segunda etapa del proceso de análisis de datos CRISP-DM, donde nuestro conocimiento sobre los conjuntos de datos aumenta y se trata la fiabilidad de los mismos, la corrección y la adaptación a los formatos necesarios.
- **Preparación de los datos:** Fase del desarrollo que tiene varias etapas comunes de los distintos métodos utilizando los conceptos de cada uno que benefician a la mejora final de los datos, preparación de los datos en la metodología de proyectos y proceso de análisis de datos CRISP-DM y la fase en el framework OSEMN, limpieza y normalizar. Fase enfocada en la limpieza, transformación, adaptación, restructuración de los datos para más tarde utilizarlos.
- **Hipótesis y modelado:** Fase del desarrollo relacionada con la etapa que se le atribuye el mismo nombre en la metodología de proyectos de ciencias de datos, apoyándose en las etapas de exploración y modelo en el framework OSEMN y la de modelo en el proceso CRISP-DM. Formulanción de distintas preguntas gracias a la exploración sobre los datos recogidos con las mejoras de readaptación y entrenando modelos con estos mismos para recuperar información de los resultados.
- **Evaluación e interpretación:** Fase del desarrollo relacionada con la etapa que se le atribuye el mismo nombre en la metodología de proyectos de ciencias de datos y apoyándose mediante las fases de interpretación del framework OSEMN y de evaluación del proceso CRISP-DM. A partir del estudio de los modelos obtenidos y la comparación de rendimientos se seleccionaran las mejores opciones.
- **Optimización:** Fase del desarrollo relacionada con la última etapa de la metodología de proyectos de ciencias de datos focalizada en la readaptación de los conjuntos de datos en busca del perfeccionamiento de los resultados.
- **Iteración:** Fase del desarrollo atemporal relacionada con la etapa añadida al framework OSEMN y a la metodología SCRUM, para el refinamiento del trabajo realizado en el proyecto.

CAPÍTULO 4

Caso de uso: Empresa Torcal Formación

4.1. Introducción

A la hora de plantear este trabajo fin de grado (TFG) como un proyecto de ciencias de datos se contemplaban utilizar datos teóricos y prooporcionados por una plataforma distribuidora de datos realizando un estudio en un entorno artificial o contactar con una empresa, administración u organización y realizar un estudio en un entorno real.

Pensamos que ejecutar este proyecto de fin de grado con información real traería consigo una serie de ventajas superiores, aunque en el camino las dificultades fueran también mayores por tener que hacer un proceso inicial de obtención de datos, limpieza, normalización que al ser una situación real conlleva bastante trabajo y complicaciones.

También encontrar un contacto aunque sea leve con un usuario real produce practicar una serie de mecanismos socio-profesionales beneficiosos de cara al futuro profesional del alumno.

La empresa seleccionada para colaborar ha sido Torcal Innovación y Seguridad S.L con CIF B29555703, más conocida como Torcal formación [18].

Figura 4.1: Logo Torcal Formación

Empresa enfocada en la formativa vial y profesional durante más de 30 años, siendo líderes en el sector, ofreciendo servicios como especialistas en formación del sector transporte (terrestre, marítimo...) maquinaría, seguridad laboral, etc. Dispone más de cuarenta certificados de profesionalidad homologados por el Servicio Público de Empleo Estatal (SEPE).

4.2. Motivación de la elección

Nuestra selección viene fudamentada por cuatro aspectos:

- Excelencia
  - Galardonada reiteradamente con el Premio Andaluz a la Excelencia.
  - Teniendo en 2019 el volumen de aprobados más alto de España.
  - Primera autoescuela en obtener certificaciones ISO-9001/2 e ISO-14001.
  - Expansión a nivel nacional.
- Innovación tecnológica
  - Test gratuitos por Internet y simulador de conducción (en sus inicios).
  - Simulador de vuelcos.
  - Primera plataforma on-line para parte teórica en los permisos de conducción.
  - Cursos en modalidad e-learning.
  - Aplicación de reservas para clases.
- Social
  - Fundación Torcal, concienciación de la educación y la seguridad vial.
  - Premio a la empresa socialmente responsable.
  - Política medio ambiental.
- Volumen de datos
  - Tiempo de vida superior a 30 años.
  - Numerosas sucursales a nivel nacional.

Como se muestra, aspectos que se identifican a la perfección con Torcal formación.

4.3. Motivación de la empresa

A continuación mostramos también el motivo por el que la empresa está interesada en un desarrollo como el mostrado en este TFG

- Mejorar procesos internos.
- Ofrecer servicios de mejor calidad.
- Ofrecer servicios más personalizados.
- Incrementar su filosofía de innovación.
- Mejorar la calidad de sus recursos tecnológicos.
- Abrir la posibilidad a futuros proyectos de I+D.

- Aumentar productividad y eficiencia.
- Captar cualquier anomalía en el funcionamiento del negocio.
- Guía de ruta para conseguir los objetivos propuestos.

4.4. Información sobre los datos obtenidos

Como hemos visto en los capítulos previos, un proyecto de ciencias de datos se sustenta por los datos obtenidos para su realización.

Los datos fueron proporcionados por la empresa en un archivo '.zip' cifrado con contraseña, en el cuál encontramos un archivo '.sql', script SQL de la base de datos de la empresa con los datos anonimizados siguiendo el protocolo de privacidad para el usuario, por lo que en este caso los datos obtenidos son en formado SQL.

Los datos venían solos sin manual sobre la base de datos. Al no disponer de información sobre la base de datos que contenga la información sobre el significado de cada tabla y de los atributos que alberga, ha implicado tener que hacer una primera fase de estudio y deducción bastante larga.

La base de datos está formada por 336 tablas, de las cuales 30 tablas estaban vacías de información y la información en las demás tablas se reparte de manera irregular pero en grandes cantidades, encontrando tablas rellenas de entre 500.000 y 3.000.000 de datos, además de encontrar datos que a la hora de realizar ciencia de datos brindan mucha información, encontramos fallos estructurales y en las recomendaciones de buenas prácticas a la hora de diseñar bases de datos que complican su acceso.

En el resto de esta subsección hablaremos más en profundidad sobre la base de datos, tablas, atributos y buenas prácticas.

Comenzamos con una subdivisión a *grosso modo* sobre tipos de tablas encontradas al realizar un análisis inicial sobre la base de datos ofrecida.

Figura 4.2: Subdivisión de la base de datos - A

En primer lugar encontramos una subdivisión entre tablas vacías y no vacías, figura 4.2. Aunque en las carrera de Ingeniería del Software y otras modalidades donde se trabaja

software, nos indican unas directrices a seguir para que los diseños sean lo más eficiente y consistentes posibles su ejecución siempre no se lleva a la práctica.

Encontramos que en esta base de datos hay casi un 10 % de las tablas sin ningún tipo de dato, gastando espacio de memoría y ralentizando el manejo general de la base de datos.

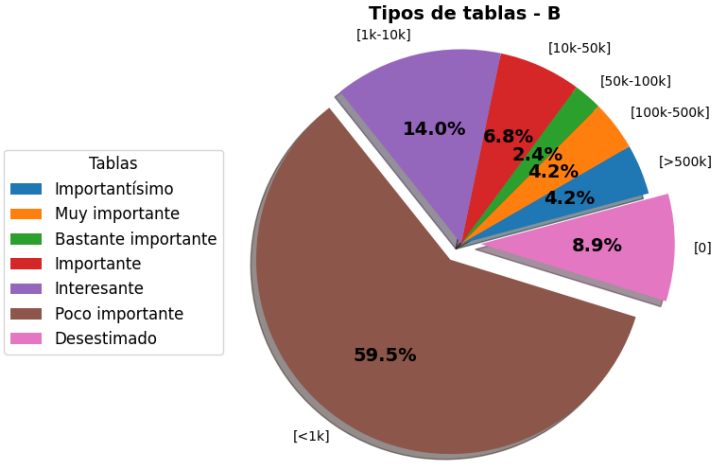

Figura 4.3: Subdivisión de la base de datos - B

La siguiente subdivisión sobre las tablas obtenidas es referido al número de datos que acumula cada tabla, figura 4.3. Cómo ya comentamos con anterioridad un estudio de los datos está totalmente relacionado con la información obtenida. Dependiendo de los datos obtenidos tendremos un estudio productivo o no.

Primero hacemos una división cuantitativa de los datos y luego haremos una subdivisión cualitativa según la exploración de la base de datos. Esta subdivisión cuantitiva nos termina dejando 7 categorías:

- **Importantísimo:** Tablas que superan las 500.000 filas.
- **Muy importante:** Tablas que su contenido se comprende entre 100.000 y 500.000 filas.
- **Bastante importante:** Tablas que su contenido se comprende entre 50.000 y 100.000 filas.
- **Importante:** Tablas que su contenido comprende entre 10.000 y 50.000 filas.
- **Interesante:** Tablas que su contenido comprende entre 1.000 y 10.000 filas
- **Poco importante:** Tablas que su contenido es menor de 1000 filas.
- **Desestimado:** Tablas que su contenido es 0.

Esta categorización se basa en el número de tablas que almacenan las tablas, siendo la categorización la prioridad de análisis de las tablas propiamnete dichas. La categoría Interesante y Poco importante tienen un rol muy secundario en la exploración de las tablas debido a su pequeño contenido de datos, pero no se desestima su importancia del todo

debido a que cuando estas tablas están relacionadas con tablas de las categorías superiores cogerán relevancia.

En relación a categorías cualitativas en las que se divide la base de datos. Encontramos 4 subdivisiones generales.

- **Administración:** Tablas relacionadas con los aspectos del funcionamiento de la empresa a nivel interno y tablas necesarías para recoger información relativa a otras.
- **Negocio:** Tablas relacionadas con los aspectos del negocio que pueden afectar en su crecimiento, el funcionamiento de la empresa a nivel externo (productos, publicidad, ofertas..
- **Alumnos:** Categoría en la que nos vamos a focalizar durante este estudios. Tablas que están estrechamente relacionadas únicamente con los alumnos.
- **Vacías:** Tablas desestimadas por no contener ninguna información de la cuál inferir resultados.

A continuación mostraremos unas distribuciones sobre aspectos de la base de datos que pueden ser de interés para analizar la estructura de la base de datos.

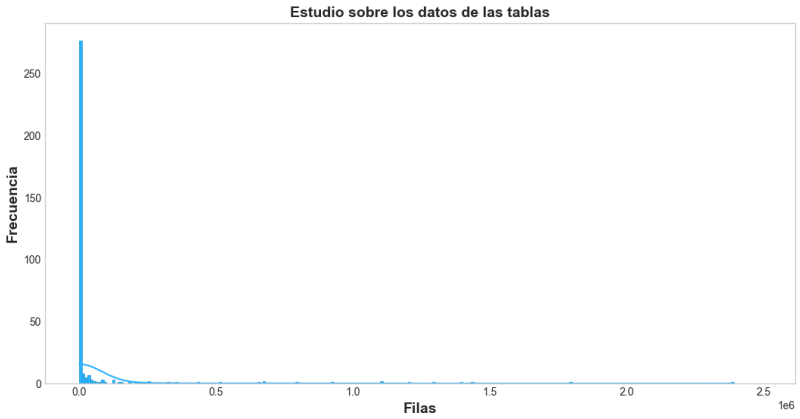

Figura 4.4: Histograma sobre el contenido de las tablas subdividido en intervalos de 10.000 filas.

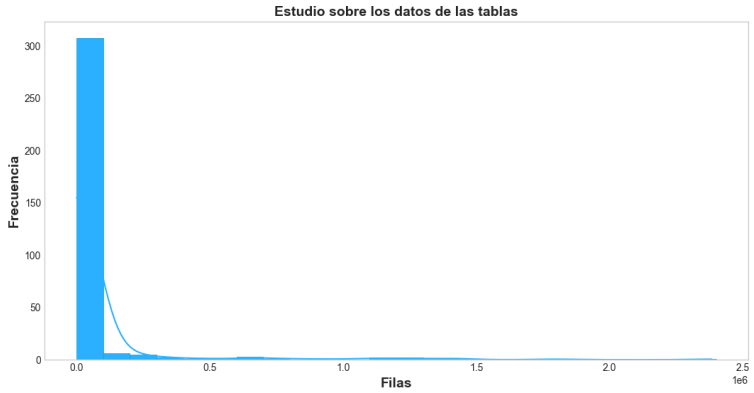

Figura 4.5: Histograma sobre el contenido de las tablas subdividido en intervalos de 100.000 filas.

En la figura 4.4 encontramos un histograma en el que las clases se dividen en rangos de 10.000 filas hasta llegar a 2.500.000, donde se infiere que la subdivisión inicial que hicimos coincide en su protuberancia sobre las primeras clases.

En la figura 4.5 las clases se dividen en rangos de 100.000 filas, se decidió mostrar esta distribución en intervalos mayores para que se visualizará mejor esta diferencia.

En otro orden de cosas, los atributos son la otra línea de información relevante sobre la base de datos y sus tablas. Comenzaremos con un estudio de la cantidad de atributos que encontramos en cada tabla y más tarde nos enfocaremos en cada atributo individual.

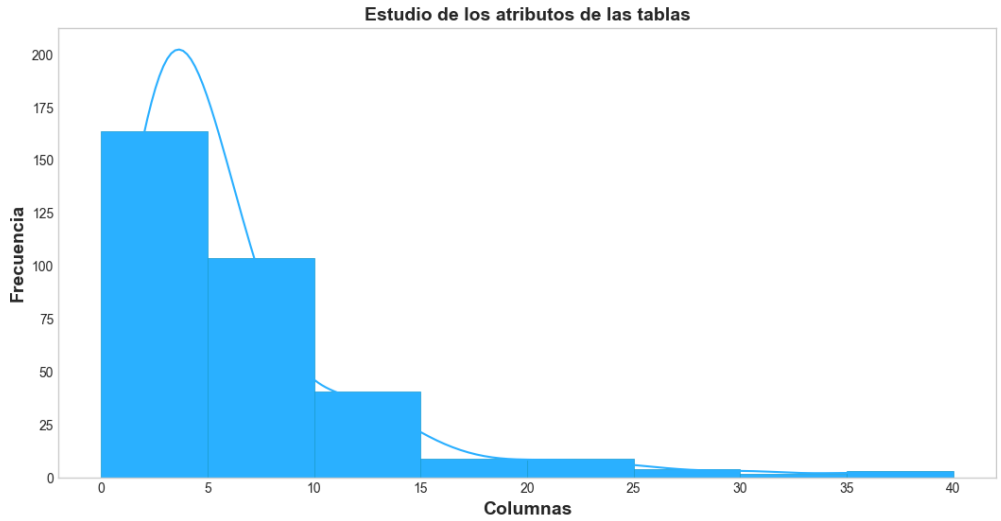

Figura 4.6: Histograma sobre los atributos de las tablas.

La figura 4.6 se encuentra acotada por los valores obtenidos en la exploración sobre las columnas, mostrando un decrecimiento claro de cara al aumento en atributos de cada tabla, algo lógico y a la vez puede ser contraproducente debido a que esa superioridad tan marcada al inicio denota una sobrecreación de tablas en muchos casos cuando la información se podía recoger en las tablas claves. Al igual que la creación de muchos campos que no se usan o que su información no es relevante puede afectar a nivel estructural y funcional en la eficiencia del modelo.

Para finalizar con las distribuciones de datos analizaremos características individuales de los atributos, con lo que se muestra alguna de las debilidades de la base de datos proporcionada.

En la figura 4.7 se encuentra uno de los problemas estructurales más notorios de la base de datos, la posibilidad de nulidad de una gran parte de los atributos muchos de ellos con un carácter de importancia estadística relevante, produce una gran serie de filas con valores nulos que desvirtuan mucho los datos en general.

En la figura 4.8 se hace un estudio sobre los tipos de datos que agrupa la base de datos, esto nos devuelve una clara focalización en datos cuantitativos sobre los valores de carácter cualitativo notable, una práctica que agiliza mucho el trabajo con la información ya que el manejo de datos numéricos es mucho más veloz que el tratamiento de los cualitativos.

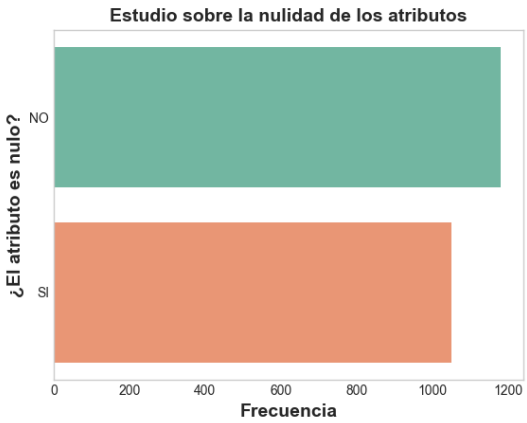

Figura 4.7: Diagrama de sectores sobre la nulidad de los datos de la base de datos.

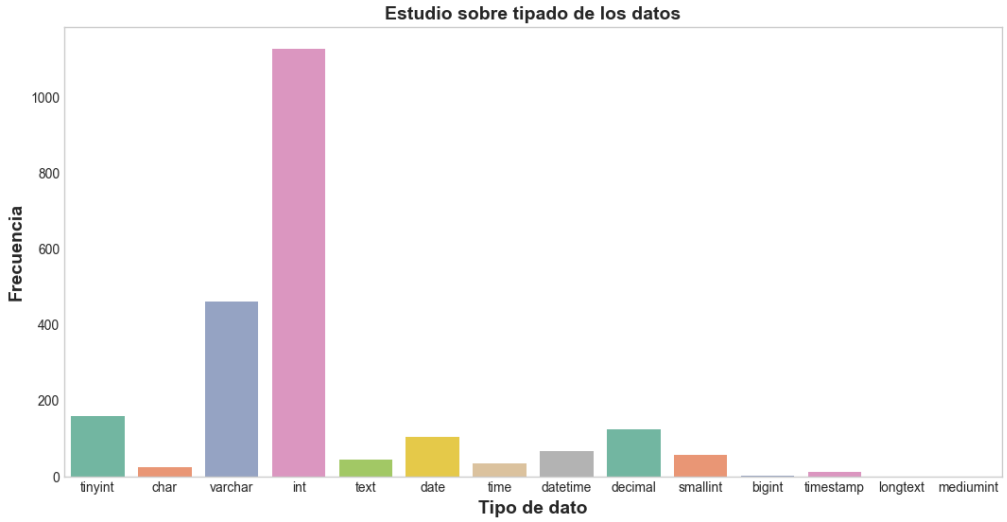

Figura 4.8: Diagrama de barras sobre el tipado de los datos de la base de datos.

A modo de resumen, los errores más comunes encontrados en la base de datos con respecto a las buenas prácticas en el diseño de bases de datos:

- Columnas completamente nulas.
- Tablas completamente vacías.
- Muchos campos nulos o vacías repartidos por la base de datos.
- Atributos sin utilidad.
- Atributos que a partir de una fecha se dejaron de rellenar pero que continuan en la base de datos.
- El patrón de los nombres es confuso, mezcla de idiomas, nombres muy similares con diferenciaciones en una letra de la que es muy dificil diferir el sentido, similitud entre nombres que complica la diferencia de uso entre ambos.

CAPÍTULO 5

Tecnologías y herramientas usadas para este TFG

5.1. Sistema operativo

El sistema será Windows por la familiarización con el sistema, pero las tecnologías utilizadas en su mayoría son adaptables a cualquier sistema operativo.

Figura 5.1: Icono Windows

Herramienta exacta: Windows 10 Home.

5.2. Base de datos

La base de datos utilizada al tener un script de sql, se tuvo que elegir una de las opciones entre SQL Server y SQL Workbench, al estar familiarizados con la segunda opción y también porque ofrece una serie de opciones con una mejor detección e interfaz fue la elegida.

SQL Workbench es una herramienta de diseño de bases de datos muy visual que integra el desarrollo, administración, creación, diseño y mantenimiento de bases de datos en SQL en un solo entorno de desarrollo integrado para el sistema de bases de datos MySQL.

Herramienta exacta: MySQL Workbench 8.0 CE

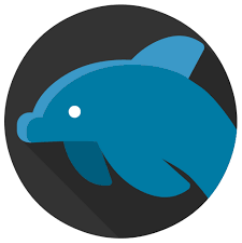

Figura 5.2: Icono MySQL Workbench

5.3. Estructura estandarizada de carpetas

Una cosa que nos ha enseñado Ingeniería del Software es que el código cuenta y la calidad de este aún más. Usualmente en ciencias de datos nos preocupa mucho los productos finales (informes de resultados, visualizaciones, conocimientos retribuidos, etc) pero la calidad del código que los genera es un aspecto importante.

Hablamos de corrección, modularización y reproducibilidad de este, los proyectos de ciencias de datos que consiguen una gran repercusión se caracterizan por tener un análisis minucioso, entrenamientos de modelos exhaustivos, visualizaciones detalladas y presentaciones estelares. La materia prima que genera todo debe estar a la altura.

Aportándonos beneficios colaborativos como personales.

- Colaborar en los análisis facilmente.
- Aprender de nuestro análisis desde el proceso y el dominio.
- Sentirse seguros en las conclusiones de donde llega nuestro análisis.
- Búsqueda de cada archivo con facilidad.

La estructura de referencia utilizada, Cookiecutter, está creada para proyectos que utilizan Python como lenguaje de programación pero no está enfocado solo a este tipo de proyectos, solamente habrá que eliminar algunos archivos.

Estandarizar un método, una estructura, una sucesión de pasos tiene una serie de objetivos esenciales en el éxito del mismo. Productividad, ganar tiempo, generalización, comprensión tanto interna como extenra, profesionalizar aún más la práctica.

Estructura base:

Herramienta exacta: Cookiecutter Data Science - cookiecutter Python package.

```
├── LICENSE
├── Makefile           <- Makefile with commands like `make data` or `make train`
├── README.md          <- The top-level README for developers using this project.
├── data
│   ├── external       <- Data from third party sources.
│   ├── interim        <- Intermediate data that has been transformed.
│   ├── processed      <- The final, canonical data sets for modeling.
│   └── raw            <- The original, immutable data dump.
│
├── docs               <- A default Sphinx project; see sphinx-doc.org for details
│
├── models             <- Trained and serialized models, model predictions, or model summaries
│
├── notebooks          <- Jupyter notebooks. Naming convention is a number (for ordering),
│                         the creator's initials, and a short `-` delimited description, e.g.
│                         `1.0-jqp-initial-data-exploration`.
│
├── references         <- Data dictionaries, manuals, and all other explanatory materials.
│
├── reports            <- Generated analysis as HTML, PDF, LaTeX, etc.
│   └── figures        <- Generated graphics and figures to be used in reporting
│
├── requirements.txt   <- The requirements file for reproducing the analysis environment, e.g.
│                         generated with `pip freeze > requirements.txt`
│
├── setup.py           <- Make this project pip installable with `pip install -e`
├── src                <- Source code for use in this project.
│   ├── __init__.py    <- Makes src a Python module
│   │
│   ├── data           <- Scripts to download or generate data
│   │   └── make_dataset.py
│   │
│   ├── features       <- Scripts to turn raw data into features for modeling
│   │   └── build_features.py
│   │
│   ├── models         <- Scripts to train models and then use trained models to make
│   │   │                 predictions
│   │   ├── predict_model.py
│   │   └── train_model.py
│   │
│   └── visualization  <- Scripts to create exploratory and results oriented visualizations
│       └── visualize.py
│
└── tox.ini            <- tox file with settings for running tox; see tox.testrun.org
```

Figura 5.3: Plantilla de proyectos Cookiecutter

5.4. Entorno de desarrollo integrado

Para el desarrollo de nuestro proyecto de ciencias de datos es necesario un entorno de desarrollo donde poder codificar, probar y desplegar soluciones de inteligencia artificial y las distintas funciones sobre visualización, limpieza y normalización de los datos.

Figura 5.4: Logo Visual Studio Code

El entorno principal utilizado será Visual Studio Code (VSCode), entorno que soporta la mayoría de lenguajes de programación existentes, una potente serie de herramientas tanto para el código, como para debugear y de despliegue. Además es un entorno con el que estamos familiarizados por el uso en distintos proyectos anteriores.

VSCode ofrece frameworks de autocompletados para lenguajes de programación específicos, herramientas de visualización para formatos '.xml', '.json' y '.csv', manejo de varias terminales diferentes desde una misma herramienta.

Herramienta exacta: Visual Studio Code 1.51

5.5. Lenguajes de programación

Los lenguajes de programación en el ámbito de la ciencias de datos más extendidos son R y Python. Ambas opciones tienen una infinidad de funcionalidades, bibliotecas con todo tipo de funciones y técnicas para realizar cada paso en el proceso de crear un proyecto de ciencias de datos.

Figura 5.5: Logo Python

Antes de seleccionar una de las dos opciones hicimos un estudio de las dos tecnologías, probando con proyectos de prueba en el proceso de aprendizaje de las herramientas de ciencias de datos. La elección fue Python. debido a que ofrecia más flexibilidad en el manejo de los datos, aumentaba la experiencia en lenguajes de programación multifuncionales, la documentación es más extensa y da más posibilidades de cara a las distintas funciones en Inteligencia Artificial.

Herramienta exacta: Python 3.8.6

5.6. Distribución de software de python

Python tiene una lista de distribuciones extensa CPython, Anaconda Python, ActivePython, PyPy, IronPython, WinPython, Python portable... En nuestro caso un proyecto de ciencia de datos, lo habitual es utilizar Python en un entorno de desarrollo directamente o utilizar la distribución Anaconda.

Figura 5.6: Logo Anaconda

Anaconda es una distribución de software o suite de código abierto enfocada en desarrolladores de Python, aunque se puede preparar el entorno para trabajar con R, donde

se recogen una serie de aplicaciones, librerías y conceptos que regularmente se usan en proyectos de ciencia de datos. Funciona como gestor de entorno, gestor de paquetes y posee una colección de más de 720 paquetes de código abierto.

Utilizaremos esta distribución como recurso al que recurrir si el entorno instalado y configurado de Python en el entorno de desarrollo fallara tener un apoyo rápido y consistente donde trabajar.

Herramienta exacta: Conda 4.8.3

5.7. Entornos virtuales

Los objetivos de los entornos virtuales es crear entornos personalizados para proyectos. Cada proyecto tiene sus propias dependencias sin importar que dependencias tienen los otros proyectos. Esta práctica es vital en proyectos de desarrollo.

Adaptas el entorno a las necesidades del proyecto en cuestión, utilizando un archivo "setup´´ donde recoges todos los procesos que debes realizar para tener el entorno adaptado a tu proyecto y un archivo "requirements´´ donde tienes registrados todas las dependencias.

Figura 5.7: Logo Pipenv

Utilizaremos los entornos virtuales que nos ofrece Python y además los entornos ofrecidos por Anaconda si fuera necesario. Pero no usaremos los entornos virtuales tradicionales sino el recurso de Python para manejar dependencias de paquetes "pipenv´´ que ofrece beneficios.

El primer beneficio reducir la necesidad de utilizar distintas bibliotecas para trabajar "pip´´ para instalación de paquetes, otra para crear el entorno, otra para manejarlo y otra que nos asocie. El segundo beneficio es manejar las interdependencias complejas de manera más cómoda. Otros beneficios son mejor características de seguridad, el grafo de dependencias en un formato más legible.

"Pipfile´´ y "Pipfile.lock´´ recogen todas las dependencias del proyecto y las interdependencias del mismo. Su legibilidad y su estructura facilita mucho el uso de estos archivos entre los distintos componentes del proyecto.

Herramienta exacta: pipenv 2020.11.15

5.8. Bibliotecas y paquetes

Utilizaremos una serie de paquetes enfocados a la ciencia de datos necesario para realizar cada fase del análisis de datos. Los paquetes se instalaran de manera local en entornos virtuales divididos por proyectos recogiendo toda la información en los archivos de control

de dependencias. En Anaconda encontraremos la mayoría de ellos y da la posibilidad de instalar otros paquetes que no estén pre-instalados.

Pandas

Biblioteca utilizada para el manejo de datos: leer archivos o bases de datos y hacer operaciones entre las columnas y los tipos de variables obtenidas (ordenar, agrupar, dividir, pivotar, totalizar).

También permite detectar valores nulos, outliers, duplicados.

Herramienta exacta: pandas 1.1.4

Pandasql

Biblioteca utilizada como lector de bases de datos para obtener resultado de querys, tablas o manejo de la base de datos.

Herramienta exacta: pandasql 0.7.3

SQLAlchemy

Biblioteca utilizada para conectar a la base de datos MySQL. Adicionalmente permite acceder a los datos siguiendo un esquema ORM.

Herramienta exacta: SQLAlchemy 1.3.20

Numpy

Biblioteca estándar de Python utilizada para operar con datos numérico. Permitiendo crear todo tipo de estructuras numéricas, múltiples dimensiones, permite transformarlas operar aritméticamente, filtrar, números aleatorios, etc.

Tuvimos problema con esta biblioteca debido a que la nueva versión 1.19.4 causa muchas incomptabilidades con Windows y tuvimos que usar la versíon anterior que era compatible con las otras bibliotecas.

Herramienta exacta: numpy 1.19.3

Matplotlib

Biblioteca utilizada para realizar el análisis exploratorio o estudiar los resultados obtenidos, produciendo gráficas informativas visuales y bastante rápidas. Su punto más débil es que los gráficos son bastante básicos.

Herramienta exacta: matplotlib 3.3.3

Seaborn

Biblioteca utilizada para el mismo propósito que Matplotlib, pero es una biblioteca embellecedora por lo que expande mucho el alcance de matplotlib dando la posibilidad al usuario de buscar gráficos más complejos si fuera neceario.

Herramienta exacta: seaborn 0.11.0

Scipy

Biblioteca enfocada a módulios de álgebra lineal, integración, optimización y estadística. Ofreciendo rutinas númericas eficientes como la optimización numérica, integración y otras en submódulos.

Herramienta exacta: scipy 1.5.4

Sklearn

Biblioteca enfocada a la Inteligencia Artificial en concreto aprendizaje máquina. Se utiliza para manejar las tareas generales en aprendizaje máquina y minería de datos como clustering, regresiones, selección del modelo, reducción de dimensión y clasificación.

Herramienta exacta: sklearn 0.0

5.9. Cuaderno de clase

Los proyectos de ciencias de datos deben ir documentados relacionando el código que se realiza con comentarios indicando y expresando los motivos de decidir hacer cualquier acción a los datos y la interpretación de los resultados obtenidos. Jupyter es la opción elegida por soportar los lenguajes que usaremos, es un entorno de desarrollo interactivo basado en web para el cuaderno de Jupyter, código y datos.

Figura 5.8: Logo Jupyter

Jupyter es un entorno flexible donde se puede configurar y preparar la interfaz de usuario para apoyar un gran rango de flujos de trabajos en ciencias de datos, computación científica y aprendizaje máquina. También es extensible y modular, se pueden escribir plugins y añadir componentes e integrarlos con ya existentes.

Esta tecnología también está soportada en documentos '.ipynb', donde desde el entorno desarrollo individual utilizando la sintaxis de formato de texto plano Markdown, desarrollamos los procesos de trabajo en ciencias de datos.

Herramienta exacta: Jupyter-notebook: 6.0.3

5.10. Control de versiones

Para la tarea de control de versiones del desarrollo del proyecto se ha utilizado Git y GitHub. La utilización de esta herramienta permite almacenar tu proyecto en la nube

donde permite que sea compartido, herramienta muy útil en entornos colaborativos. Permite guardar los distintas cambios que se realizan y visualizar si se han hecho, donde se han efectuados y poder volver a versiones anteriores del proyecto para recuperar la información en cuestión.

Figura 5.9: Logo Git

Muy útil para visualizar el progreso del desarrollo, tener un registro de incidencia y asegurarse la estabilidad del proyecto. Pudiendo volver a versiones anteriores con facilidad y ofreciendo una gran comodidad para el trabajo multiplataforma y colaborativo.

Figura 5.10: Logo Github

Herramienta exacta: git 2.29.2.windows.2

5.11. Herramientas de apoyo

Se han utilizado algunas herramientas alternativas para asegurarse que la realización del proyecto se ejecutaba en un ritmo correcto, constante, eficiente y seguro.

Trello: Una de las aplicaciones de gestión de proyectos y de productividad personal, herramienta donde se pueden aplicar las distintas metodologías de desarrollo de proyectos como de productividad personal (SCRUM y GTD).

Figura 5.11: Logo Trello

Entorno dispuesto por tableros virtuales, que se subdiden en tablones virtuales verticales relacionados con el tipo de tareas que se realizaran en ese tablón. En estos tablones se pueden ir añadiendo notas que estarían relacionadas con las distintas tareas. Además, hay una serie de funciones muy útiles como etiquetados para diferenciar los tipos de tareas, asignaciones de tareas, fechas de vencimiento.

Comunicación:

Skype: Software que permite comunicaciones de texto, voz y video sobre Internet (VoIP), la finalidad de esta herramienta es conectar a usuarios y que puedan comunicarse desde cualquier lugar. Herramienta utilizada para la comunicación directa tanto en audio como en videoconferencia.

Figura 5.12: Logo Skype

Gmail: Servicio de correo electrónico, centrado en la comunicación entre personas, organizaciones, empresas tanto para comunicaciones o envios de documentos. Herramienta utilizada para una frecuente comunicación de actualización de situación.

Figura 5.13: Logo Gmail

CAPÍTULO 6

Implementación de un proyecto real

6.1. Entendimiento del negocio

Esta primera fase es muy importante para la ejecución del proyecto ya que determinará el enfoque de los pasos siguientes, está incluido en el proceso CRISP-DM dominante en los procesos de minería de datos.

El objetivo es definir que se busca con el proyecto y las razones por las que se quiere alcanzar esta meta, está subdividido en cuatro tareas. Para esta parte conocer la empresa y el contacto con la misma es vital, en este caso representando ese *stakeholder* por los tutores.

Indentificar las metas de negocío

Después de explorar la información de la empresas, sus metas anteriores, productos, necesidades y comunicar con el representante de la empresa.

La principal atención de la empresa son sus clientes o alumnos en este caso, siendo el foco principal de la empresa. Información que puede ser clave en este aspecto puede ser la situación personal, geográfica, económica, fechas de ingreso, transacciones, etc.

Las metas del negocio será incrementar el número de clientes a nivel nacional y a nivel individual, la fidelización de los mismos (consumir varios productos de la empresa), mantener la efectividad en la superación de los productos obtenidos (al realizar exámenes).

El criterio que se tomará para evaluar el éxito de los resultados será una ganancia sobre la información de los clientes que guíe las medidas a tomar ratificado por el representante de la empresa.

Evaluar la situación

Los recursos de los que obtenemos son una base de datos sobre el negocio al completo en crudo, un estudiante cómo científico de datos apoyado por dos profesores cómo científicos de datos expertos, un ordenador de trabajo y distintas herramientas software.

Todo el proceso de manejo de datos y de estudio siguen los prótocolos de seguridad y consistencia corroborado por la empresa.

Los posibles problemas y riesgos para la realización del proyecto se aseguran con una estructura en la nube que asegura una copia protegida de problemas de conexión, batería o perdidas de información. Los problemas médicos e interpersonales de los integrantes del proyecto están contemplados en el tiempo de la realización del proyecto.

Definir las metas de la minería de datos

Los objetivos a nivel de minería de datos son conseguir una lista de diagramas y visualizaciones que nos permita aumentar ese conocimiento sobre los clientes.

En esto se incluyen cálculos estadísticos, distribuciones, diagramas, agrupaciones de los datos y un apartado de predicciones.

La presentación de diagramas que retribuyan información sobre los clientes, aplicación de algoritmos de agrupación y alguna técnica de predicción con éxito.

Producir el plan de proyecto

Este apartado está presentado en las siguientes secciones de este capítulo donde se aplica el plan de proyecto para poder obtener los resultados buscados.

6.2. Adquisición de los datos

Primera fase del ciclo de vida de un proyecto de ciencia de datos, esta fase trata de la obtención de los datos necesarios para poder realizar el proyecto de ciencia de datos. Hay distintas maneras de adquirir los datos información externamente mediante llamadas a API o uso de *web scraping* o internamente recuperando los datos desde archivos de csv simple en local o una base de datos extensa desde almacenamiento de datos. Tener conocimientos sobre procesos de ETL y lenguaje de consultas son útiles para esta fase.

En nuestro caso el conjunto de datos está recogido en una base de datos MySql explicada más detenidamente en la Sección 4.3. Los pasos para la recolecicón han sido utilizar consultas extensas para recuperar los datos que tengan valores válidos, luego pasar estos datos desde los archivos ".sql´´ a ".csv´´ para manejar los datos con las herramientas de ciencia de datos ofrecidas por python.

Para esta transformación de tipos de archivos hay una opción de exportar los archivos desde MySQL Workbench pero también python nos ofrece bibliotecas para hacer este proceso. Haremos uso de la biblioteca *sqlalchemy* también se puede usar la biblioteca *pandasql*.

Listing 6.1: Código Python para recuperar los datos y transformar el tipo de archivo

```
import pandas as pd
from sqlalchemy import create_engine

engine = create_engine('mysql://user:passworda@host/
nombreDeLaBaseDeDatos?', encoding=``Tipo de dato    )

data = pd.read_sql_table(Nombre de la base de datos, engine)
data = pd.read_sql_query(Consulta de MySql, engine)

data.to_csv(``path del archivo exacto    )
```

Listing 6.2: Consulta en SQL para recuperar conjunto de datos específico y con valores válidos

```
# QUERY WITH INFO ABOUT MONEY TOTAL/FIRST PAYED DATE/NTYPES OF PAYS
 SELECT C.alumatc_pk, C.alumat_alumno_matricula_alumat_pk,
 C.alumatc_v_numprovincial, C.alumatc_v_cp, C.alumatc_v_direccion,
 C.alumatc_v_provincia, M.alumat_falta ,C.alumatc_c_fnacimiento,
 C.alumatc_c_cp, C.alumatc_f_direccion, C.alumatc_f_provincia,
 C.alumatc_f_cp, A.nie_nivel_estudios_nie_pk, N.nie_desc,
 A.sil_situacion_laboral_sil_pk, S.sil_desc,
 A.alu_permite_email, A.alu_permite_sms,
 P.per_sexo,
 SUM(TRAN.tran_importe_pvp) as totalPagado,
 COUNT(DISTINCT TRAN.fp_forma_pago_fp_pk) as nDistFormasPago,
 TRAN.tran_fechahora as fechaPrimerPago
 FROM torcal_erp_dump.alumatc_cabecera_contrato C
 INNER JOIN torcal_erp_dump.alumat_alumno_matricula M
        ON M.alumat_pk = C.alumat_alumno_matricula_alumat_pk
 INNER JOIN torcal_erp_dump.alu_alumno A
        ON M.alu_pk = A.alu_pk
 LEFT JOIN torcal_erp_dump.nie_nivel_estudios N
        ON A.nie_nivel_estudios_nie_pk = N.nie_pk
 LEFT JOIN torcal_erp_dump.sil_situacion_laboral S
        ON A.sil_situacion_laboral_sil_pk = S.sil_pk
 INNER JOIN torcal_erp_dump.per_persona P
        ON A.per_persona_per_pk = P.per_pk
 INNER JOIN torcal_erp_dump.tran_transacciones TRAN
        ON M.alumat_pk = TRAN.alumat_alumno_matricula_alumat_pk
 WHERE alumatc_c_fnacimiento >=    1900   -00-00
 AND  alumatc_c_fnacimiento <=    2006   -00-00
 AND MONTH(alumatc_c_fnacimiento) != 0 AND
 (YEAR(alumat_falta) - YEAR(alumatc_c_fnacimiento)) >= 15
 AND (YEAR(alumat_falta) - YEAR(alumatc_c_fnacimiento)) < 80
 AND P.per_sexo < 3 AND alumatc_c_cp <> ''
 AND length(alumatc_c_cp) = 5 AND alumatc_c_cp >= '01000'
 AND alumatc_c_cp <= '05000'
        group by alumat_alumno_matricula_alumat_pk
```

```
having totalPagado > 0;
```

6.3. Entendimiento de los datos

Esta tercera fase está relacionada con la segunda fase del proceso CRIPS-DM, donde se exploran los datos para comprobar si los datos obtenidos son apropiados para nuestras necesidades, en base a los datos que obtenemos se podrá replantear metas y los pasos a seguir para alcanzar los objetivos. Esta fase también se divide en cuatro tareas:

Recopilación de los datos

Esta tarea se realizó en la fase anterior, verfificando el acceso a los datos y el tipo recurso a utilizar.

Al recopilar los datos en la fase anterior ya teniamos planteados en la primera fase los tipos de datos que necesitariamos: transacciones, fechas, información sobre los clientes (procedencia, estudios, edad).

Las tablas de la base de datos en la que nos hemos focalizado han sido las relacionadas con los alumnos, fue bastante complejo el tratamiento de la base de datos debido a problemas de las definiciones de tipados de los datos y los conjuntos de datos incompletos.

Descripción de los datos

Esta tarea está definida en su mayoría en la Sección 4.3, enfocada en definir los tipos de datos aportados genéricamente. La base de datos nos ha aportado una serie de campos que han sido de mucha utilidad con respecto al proyecto cumpliendo con las expectativas a nivel de campos requeridos (fecha de nacimiento, código postal, sexo, situación laboral, nivel de estudios, fechas de acceso), el problema se encuentra en la falta de datos, tipado o errores en el rellenado de los datos.

Exploración de los datos

Esta tarea se trata de examinar los datos más exhaustivamente. En esta fase se han utilizado técnicas de estadística básica para ir comprobando la información que nos aportaba la base de datos y ver cuales datos estaban rellenados con más exactitud o con menos.

En este proceso aplicamos cálculos generales de estadística básica usando las bibliotecas *numpy*, *scipy* y *pandas*, en ellas encontramos métodos para calcular media, desviación típica, moda, varianza, cuartiles, mínimos y máximos.

También utilizamos la biblioteca *matplotlib* para realizar diagramas genéricos e iniciales para aumentar el conocimiento sobre los datos.

Verificar la calidad de los datos

En esta tarea examinamos si los datos obtenidos se adaptan correctamente a lo esperado.

Los datos que necesitamos existen pero se encuentran muchos errores de almacenamiento que han complicado el trabajo con los datos, listaremos unos pocos encontrados:

- Fechas de nacimiento erróneas produciendo edades negativas o edades superiores a la edad humana.

- Fechas de acceso erróneas, antes de que nacieran los clientes.
- Datos nulos en muchos campos.
- Códigos postales inexistentes en España.

6.4. Preparación de los datos

Esta fase del proyecto de ciencia de datos es la que más tiempo consume, más aún cuando la experiencia en proyectos de ciencia de datos es limitada.

En esta etapa nos hemos focalizado en encontrar los valores de los campos que son erróneos no solo a nivel lógico sino teórico, manejar los datos vacíos, manejar valores fuera de lo habitual o *outliers*, manejar los tipos de datos obtenidos que se adapten a tipos manejables.

Para esta parte hemos utilizado técnicas de consultas SQL para una normalización y limpieza inicial más genérica y luego la utilización de funciones de la biblioteca de python, *pandas* para eliminar filas con valores nulos, sustitución de valores en base a encontrar outliers erróneos, tipificar campos con tipos exactos.

Lista de funciones utilizadas de pandas:

- *astype()* Utilizando para convertir en el tipo decidido.
- *to_numeric()* Utilizado para transformar a números.
- *drop_duplicates()* Eliminar los valores duplicados.
- *fillna()* Rellenar los campos nulos por el valor indicado
- *isnull()* Comprobar los valores nulos del conjunto de datos.
- *isna().any()* Comprobar que columnas tienen algún valor nulo.
- *replace()* Sustituir valores por otros

6.5. Hipótesis y modelado

Esta fase del desarrollo del proyecto de datos se centra en la exploración de los datos y su información para poder recabar supuestos y aplicar distintos tipos de modelos para aplicarlos.

Al realizar los análisis estadísticos iniciales y la exposición de diagramas se veían grupos de edades, zonas geográficas y gastos por parte de los clientes que se podrían identificar mediante el uso de técnicas de machine learning.

Los modelos creados se basan en la utilización de los algoritmos de clasificación no supervisados K-Means y jerárquicos en este caso el aglomerativo. Dejando

Listing 6.3: Código Python para la aplicación de los distintos algoritmos en alto nivel

```
from sklearn.cluster import AgglomerativeClustering
from sklearn.cluster import KMeans


K–Means = KMeans(n_clusters=4).fit(df)
centroids = KMeans.cluster_centers_

aggl = AgglomerativeClustering(n_clusters=2, affinity='euclidean', linkage=
aggl.fit_predict(X)
```

En la sección de resultados 7 se detalla todos los diagramas generados para realizar las distintas cuestiones y la aplicación de los modelos a estos conjuntos de datos para llegar a conclusiones en el siguiente apartado.

6.6. Evaluación e interpretación

Esta fase del desarrollo se enfoca en la aceptación de las hipótesis generadas en el apartado anterior y la interpretación de los resultados obtenidos además de la presentación de los mismos.

La evaluación sobre los modelos aplicados se detallará también en la sección de resultados 7 con las distintas figuras y las explicaciones de las mismas. La interpretación de los mismos se detalla en esa sección, en la conclusión y además en la presentación del TFG donde se explicará más exhaustivamente.

6.7. Optimización

Esta fase del desarrollo del proyecto de datos es vital para la refinación de los resultados finales sean lo más fiables posibles.

Este paso ha sido uno de los más accedidos durante el proyecto debido a la inexperencia en proyectos de ciencia de datos. El refinamiento de los datos es una tarea que requiere iteraciones hasta encontrar los datos que entrenan correctamente a los modelos.

En nuestro caso, para entrenar a los algoritmos de agrupamiento y clasificación se tuvo que readaptar varias veces la tipifación, y aplicar funciones de la biblioteca *pandas* de *python* para mejorar el formato que se entrega a los modelos.

Por ejemplo en las fechas de nacimiento y de acceso primero utilizamos las fechas en crudo, luego utilizamos la edad en formato de días y los acceso anuales y por último las edades con número de años y accesos con el mes exacto.

6.8. Iteración

Esta es la fase más atemporal del proyecto, la búsqueda de resultados lo más válidos posible y aplicar mejores funciones o técnicas produce que sea una fase vital.

En nuestro caso el proyecto ha ido pasando entre fases dependiendo de las necesidades de cada momento. Después de encontrar un conjunto de datos interesantes y haber exprimido su conocimiento al máximo, se fue accediendo a otros conjuntos de datos que se podían combinar con estos.

Un proceso incremental que asegura resultados y unos resultados que se van refinando.

CAPÍTULO 7

Resultados

7.1. Estadística descriptiva inicial aplicada al proyecto

En esta sección analizaremos los valores estadísticos generales calculados a partir de los valores recogidos en los distintos conjuntos de datos obtenidos.

Cálculos estadísticos

A la hora de realizar los cálculos estadísticos aparecen dos problemas clave, el tipado de los datos obtenidos de las tablas mysql y la falta de valores correctos en los conjuntos de datos, valores nulos o erróneos.

Para resolver la primera cuestión se transformaran los elementos que aparecen todos en tipo *object*, en el tipo de dato que debería tener cada uno, tarea que la obtención de datos vacíos complica teniendo que realizar un paso previo para poder convertir luego cada dato en su tipo.

Para calcular los valores lo más fielmente posible a los valores reales seguimos las recomendaciones para la normalización de datos:

- Eliminar las filas con valores nulos
- Sustituir los valores por la media, moda, mediana.
- Utilizar la desviación de los valores vecinos.
- Asignar una categoría única propia.

En base a los datos obtenidos por acercanos a la fidelidad de los datos pondremos en práctica las dos primeras opciones para comprobar el impacto en la totalidad de los datos y la elección de una para el trabajo general de los mismos. En primera instancia obtendremos las modas con su frecuencia, valores únicos y el computo de valores totales obtenidos.

	alumat_falta	alumatc_c_fnacimiento	fechaPrimerPago
count	137530	137530	137528
unique	5214	17452	129074
top	2017-09-20	1991-02-11	2001-10-02 00:00:00
freq	177	39	35

Tabla 7.1: Tabla sobre modas y valores únicos - A

	nDistFormasPago	nie_nivel_estudios_nie_pk	sil_situacion_laboral_sil_pk
count	137530	115770	137521
unique	6	5	4
top	1	2	1
freq	74984	57735	70068

Tabla 7.2: Tabla sobre modas y valores únicos - B

	per_sexo	alu_permite_sms	alu_permite_email
count	137530	137530	137530
unique	2	2	2
top	1	1	1
freq	91715	132283	128211

Tabla 7.3: Tabla sobre modas y valores únicos - C

A partir de estas tablas 7.1 y 7.2 encontramos los valores de las modas llamados “top´´, su frecuencia relacionada “freq´´, el número de valors únicos del conjunto de datos específico “unique´´ y “count´´ el número de datos totales de cada conjunto de datos. El siguiente paso ha sido aplicar la primera de las opciones (eliminación) para poder

hacer los cálculos estadísticos y hallar las medidas de tendencia central de las variables con valores nulos, cómo se puede inferir de la tabla anterior tres de los conjuntos de datos tienen algunos valores nulos “fechaPrimerPago´´,“nie_nivel_estudios_nie_pk´´ y “sil_situacion_laboral_sil_pk´´.

Antes de realizar los cálculos pertinentes eliminamos las filas con valores nulos, adaptaremos los datos al tipo *integer* y renombraremos estos atributos en “nvlEstudios´´ relacionada con el nivel de estudio de los clientes y “sitLaboral´´ relacionada con la situación laboral de los clientes.

	nvlEstudios	sitLaboral
count	49512.0	49512.0
mean	1.96	2.01
std	1.02	1.31
min	1.0	1.0
25%	1.0	1.0
50%	2.0	1.0
75%	2.0	3.0
max	5.0	4.0

Tabla 7.4: Tabla de estadistica básica sobre atributos con valores nulos.

Como se puede inferir en la tabla 7.4 la media y la mediana de *nvlEstudios* utilizando la técnica de redondeo es la misma, 2.0, coincidente también con la moda hallada en las primeras tablas en la variable nie_nivel_estudios_nie_pk, variable en crudo antes de normalizarla.

Sobre la segunda variable que nos atañe *sitLaboral*, media y mediana difieren siendo la primera redondeada a 2.0 y la segunda 1.0. Es de interés el estudio de ambas opciones para luego elegir una de las dos opciones. En esta variable moda y mediana coinciden al igual que el anterior en 1.0.

En base a este análisis se calcularán las medidas de tendencia central básicas a partir de la normalización de estas dos variables (nvlEstudios y sitLaboral), tanto por eliminación como por sustitución usnado la media y la moda respectivamente. Los resultados son similares acercándose en todos los casos al valor 2.0. La elección fue trabajar con la sustitución por moda.

	edad	alumCP	pagoTotal
count	137530.0	137530.0	137530.0
mean	27.45	28925.85	604.61
std	10.22	2967.11	592.61
min	14.0	1002.0	0.01
25%	19.0	29010.0	129.0
50%	25.0	29200.0	407.1
75%	34.0	29620.0	946.85
max	79.0	50770.0	8760.4

Tabla 7.5: Tabla con cálculos sobre atributos despues del rellenado - A

	numfpago	nvlEstudios	sitLaboral
count	137530.0	137530.0	137530.0
mean	1.6	2.15	2.19
std	0.75	0.85	1.29
min	1.0	1.0	1.0
25%	1.0	2.0	1.0
50%	1.0	2.0	1.0
75%	2.0	2.0	3.0
max	6.0	5.0	4.0

Tabla 7.6: Tabla con cálculos sobre atributos despues del rellenado - B

	sexo	ynEmail	ynSMS
count	137530.0	137530.0	137530.0
mean	1.33	0.93	0.96
std	0.47	0.25	0.19
min	1.0	0.0	0.0
25%	1.0	1.0	1.0
50%	1.0	1.0	1.0
75%	2.0	1.0	1.0
max	2.0	1.0	1.0

Tabla 7.7: Tabla con cálculos sobre atributos despues del rellenado - C

7.2. Representaciones de las distribuciones de interés

En esta sección analizaremos las distintas distribuciones de datos obtenidas, para poder recabar conclusiones que sean productivas con respecto a los conjuntos de datos que son de interés de estudio.

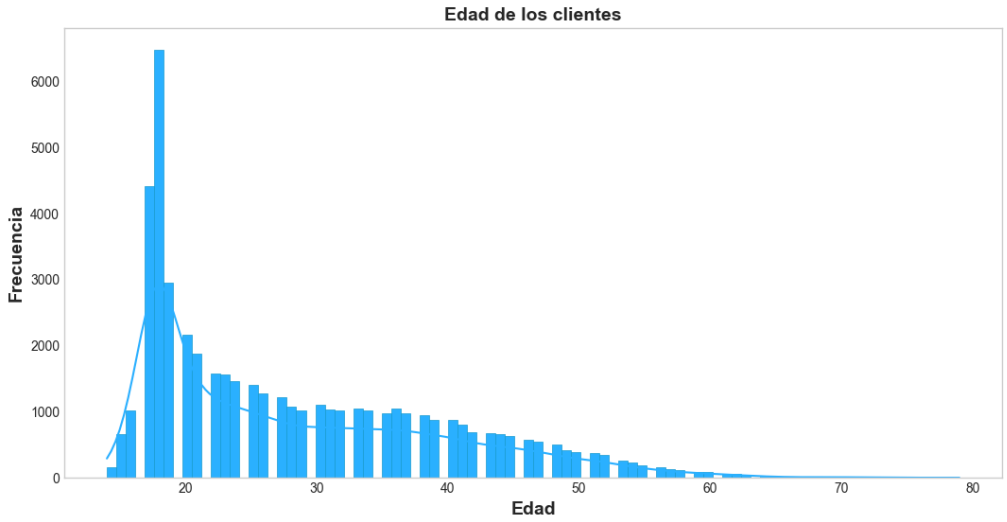

Figura 7.1: Histograma sobre las edades de los clientes y curva de ajuste

Esta gráfica es un histograma de las edades de los clientes, siendo el **eje X** las edades y el **eje Y** la frecuencia de estas edades. Dependiendo de la altura de los rectángulos indicará la frecuencia de clientes con esa edad.

En la figura 7.1 se puede inferir que la edad clave en los clientes está entre los 18-20 años siendo la edad marcada por el carnet de conducir. Después de este pico tan acentuado tras la edad de los carnets de vehículos de baja cilindrada y motocicletas (15-17 años) encontramos un decrecimiento claro hasta las personas más mayores.

Hay una estabilidad entre los 27-45 años, siendo un tramo de edad en que todos los productos que ofrece la autoescuela están disponibles para los clientes. Este gráfico se realiza de un histórico de más de 40 años en el que se visualiza que el patrón de rango de edades es constante.

La distribución que encontramos está bastante concentrada con respecto a la muestra obtenida, por lo que es de tipo Leptocúrtica a la izquierda, aunque no está tan concentrada como otras pero no lo suficientemente medio cómo para ser mesocúrtica.

Para tener una visión de la distribución de las edades más general se ha decidido mostrar un diagrama subvidivido por intervalos de 10 años en el que se ve el escalonamiento según la edad de los clientes aumenta, más adelante se agruparan los datos de las edades para fijar una relación entre los datos.

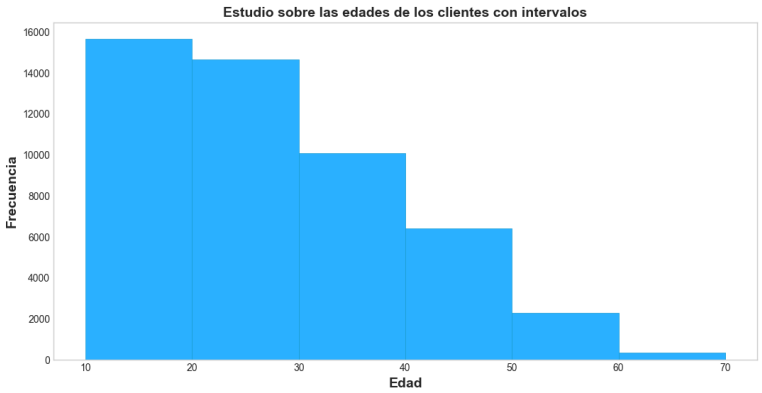

Figura 7.2: Histograma sobre las edades de los clientes en intervalos de 10 años

Pasaremos a analizar las localizaciones de los clientes, para poder revisar las situaciones geográficas de los clientes.

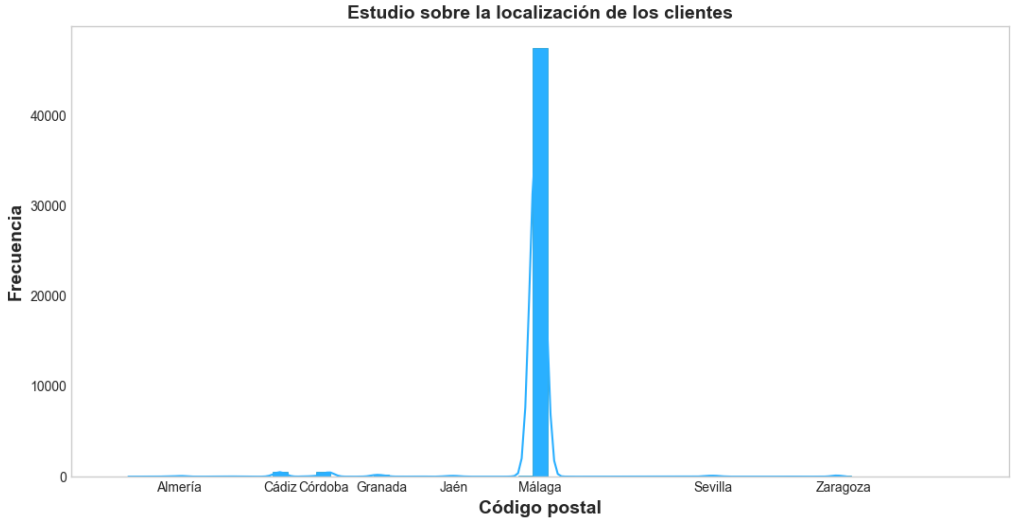

Figura 7.3: Histograma sobre la localización geográfica de los clientes

La figura 7.3 es un histograma de las localizaciones de los clientes, siendo el **eje X** los códigos postales y el **eje Y** frecuencias de estos códigos postales. Para mejorar la inferencia de conocimiento a las zonas donde hay un grupo de clientes considerable se han indicado las zonas exactas con el nombre de la provincia exacta en el **eje X**. Dependiendo de la altura de los rectángulos indicará la frecuencia de clientes que pertenecen a esa localización geográfica.

La empresa es oriunda de Málaga por lo que se denota una gran diferencia en el histórico de localizaciones de los clientes, después se ve una predominancia andaluza de los códigos postales con algunas zonas por distintas partes de España como Zaragoza.

La distribución que encontramos está muy concentrada con respecto a la muestra obtenida, por lo que es de tipo Leptocúrtica en el centro derecha.

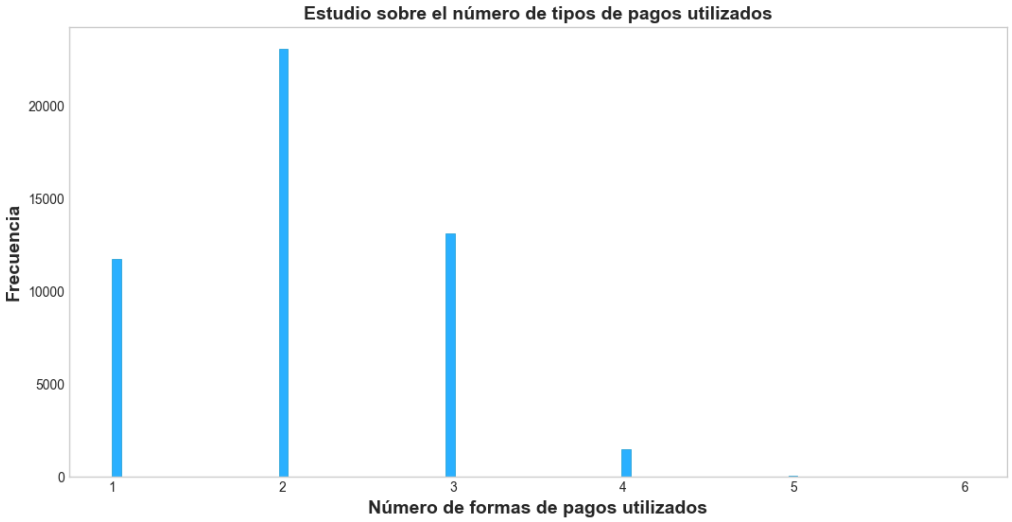

Figura 7.4: Histograma sobre el número de tipos de pagos distintos utilizados

La figura 7.4 es un histograma del número de distintos tipos de pagos que se utilizan por parte de los clientes, siendo el **eje X** los número de formas distintas de pago y el **eje Y** frecuencias de clientes que las han aplicado. Dependiendo de la altura de los rectángulos indicará la frecuencia de clientes que han utilizado un número de formas de pagos.

En este gráfico podemos inferir la predisposición de usar 2 formas de pago que normalmente son metálico y tarjeta. La fidelización de los clientes viene muy relacionada con la domiciliación de pagos por lo que los dos valores predominantes que son 2 y 3 indican ese uso de varias formas de pago que nos acerca más a esta lógica.

Figura 7.5: Histograma sobre las transacciones de los clientes

La figura 7.5 es un histograma de las transacciones realizadas por los clientes, siendo el **eje X** las transacciones en euros y el **eje Y** frecuencias de clientes que han realizado estas. Dependiendo de la altura de los rectángulos indicará la frecuencia de clientes que han realizado ese valor de transacciones.

La concentración más marcada de los pagos está entre los 0-500 euros, donde se se aglomera la mayoría de los pagos de los clientes. El decrecimiento hacía el aumento de las transacciones es claro aunque entre los 500-1200 euros hay una estabilidad de la frecuencia de clientes que realizan esa cantidad. A partir de esos valores se encuentra un grupo menor de los clientes pero eso denota qué clientes pueden consumir bastantes productos de la empresa.

La distribución que encontramos está muy concentrada con respecto a la muestra obtenida, por lo que es de tipo Leptocúrtica a la izquierda, aunque encontramos una segunda meseta más centrada alrededor de los valores 500 y 1000 euros.

Para tener una visión de la distribución de las transacciones más clara se ha decidido mostrar un histograma subvidivido por intervalos de 500 euros en el que se ve el escalonamiento según las cantidades de euros aumentan progresivamente.

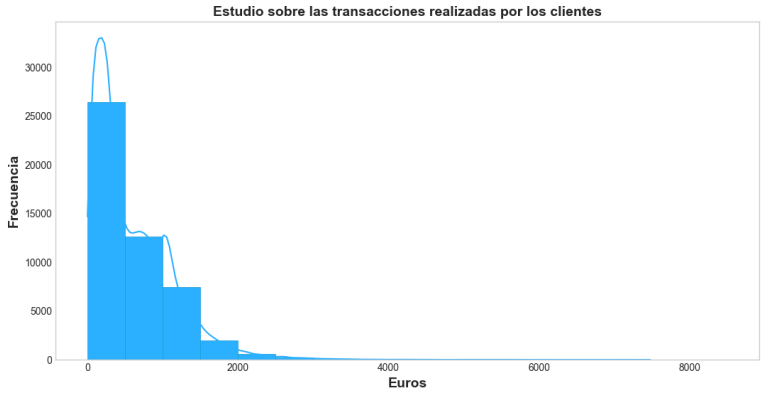

Figura 7.6: Histograma sobre las transacciones de los clientes en intervalos de 500

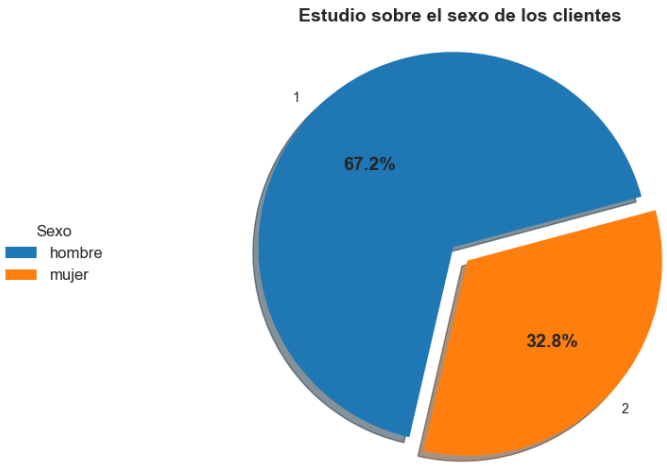

Figura 7.7: Diagrama de los sectores sobre el sexo de los clientes

Este gráfico es un diagrama de tarta sobre el sexo de los clientes, siendo dos sectores delimitados por **1 - hombre - azul** y **2 - mujer - naranja**. Dependiendo del área de los sectores indicará la frecuencia de clientes que están identificado con ese sexo.

Los porcentajes sobre hombres y mujeres están decantados para los clientes con sexo hombres, pero se han ido igualando en el progreso del tiempo. Esto solo nos indica el histórico del sexo de los clientes.

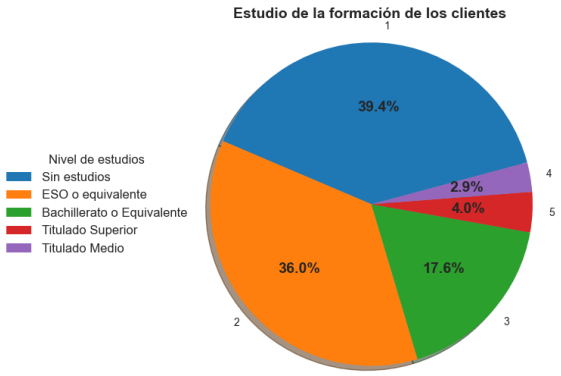

Figura 7.8: Diagrama de los sectores sobre el nivel de estudio de los clientes

Este gráfico es un diagrama de tarta sobre el nivel de formación de los clientes, siendo dos sectores delimitados por **1 - Sin estudios - azul**, **2 - ESO - naranja**, **3 - Bachillerato - Verde**, **4 - Titulado Medio - Morado** y **5 - Titulado Superior - Rojo**. Dependiendo del área de los sectores indicará la frecuencia de clientes que están relacionadas con esos niveles estudios.

Los porcentajes de titulados medios y superiores es mínimo estando totalmente relacionado con la obtención del carnet del coche. La gran mayoría de los clientes se encuentran en la categoría de sin estudios o teniendo la ESO o equivalente siendo el grupo de sin estudio sel primero, una situación que en la actualidad en el que un mayor por centaje de personas consiguen estudios se debería estar balanceando cada vez más a los estudios.

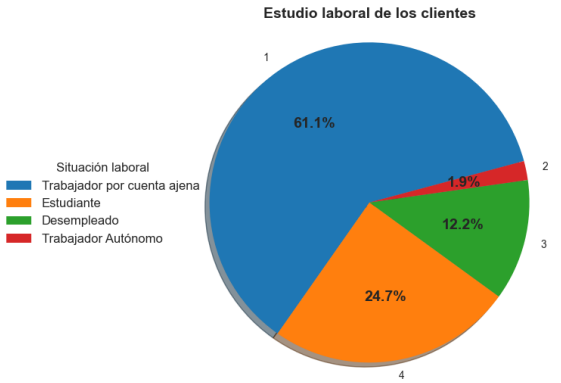

Figura 7.9: Diagrama de los sectores sobre la situación laboral

Este gráfico es un diagrama de tarta sobre la situación laboral de los clientes, siendo dos sectores delimitados por **1 - Trabajador por cuenta ajena - azul**, **2 - Trabajador Autónomo - Rojo**, **3 - Desempleado - Verde** y **4 - Estudiante - Naranja**. Dependiendo del área de los sectores indicará la frecuencia de clientes que están relacionadas con esa situación laboral.

Los porcentajes predominantes son los trabajadores por cuenta ajena y estudiantes, denotando una clara relación con la edad predominante de clientes siendo bastante temprana cómo para la creación de un negocio por cuenta propia trabajadores autónomos un 1.9 % y teniendo unos números de desempleados bastante bajos.

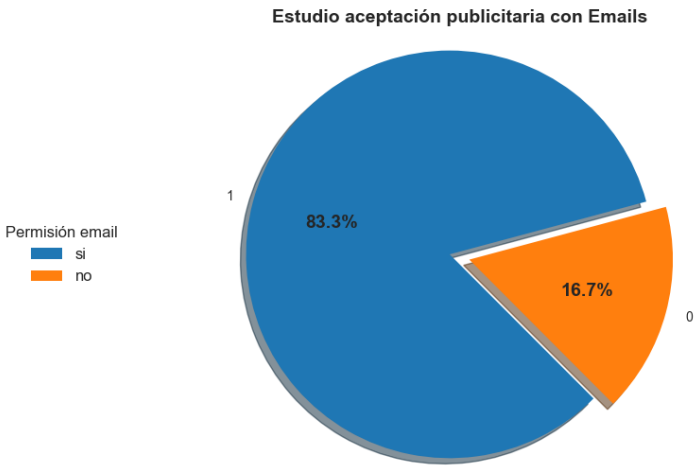

Figura 7.10: Diagrama de los sectores sobre permisos otorgados a Torcal Formación para enviar emails

Este gráfico es un diagrama de tarta sobre la permisión a emails de los clientes, siendo dos sectores delimitados por **1 - SI - azul** y **2 - NO - Rojo**. Dependiendo del área de los sectores indicará la frecuencia de clientes que están relacionadas con esa permisión de envío de información mediante sms.

Este diagrama nos muestra un conjunto de los clientes que tienen predisposición al envío de sms, por lo que son un grupo que acepta el intercambio de información mediante este medio siendo susceptibles de aceptar campañas de publicidad que puedan ser atractivas para ellos.

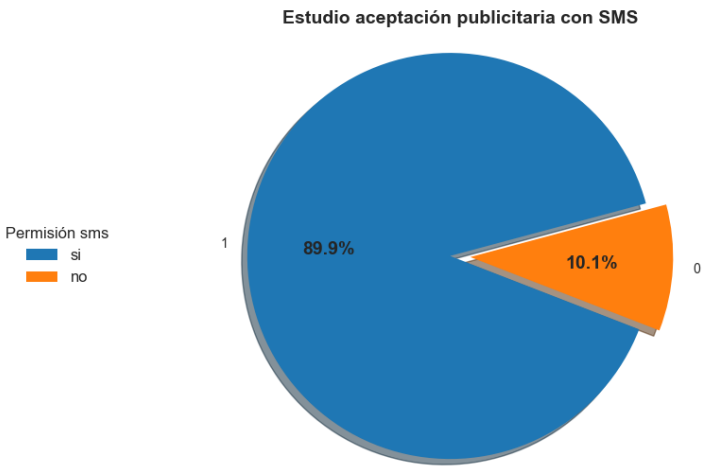

Figura 7.11: Diagrama de los sectores sobre permisos otorgados a Torcal Formación para enviar sms

Este gráfico es un diagrama de tarta sobre permisos otorgados a Torcal Formación para enviar sms de los clientes, siendo dos sectores delimitados por **1 - SI - azul** y **2 - NO - Rojo**. Dependiendo del área de los sectores indicará la frecuencia de clientes que están relacionadas con esa permisión de envío de información mediante emails.

Exactamente el mismo porcentaje que el diagrama de sectores anterior. Este diagrama nos muestra que este conjunto de los clientes relacionado con el anterior tienen predisposición al envío de email, por lo que son un grupo que acepta el intercambio de información mediante ambos medios siendo doblemente susceptibles de aceptar campañas de publicidad que puedan ser atractivas para ellos.

7.3. Correlaciones entre las variables

En esta sección nos enfocaremos en la búsqueda de patrones que relacionen variables entre sí.

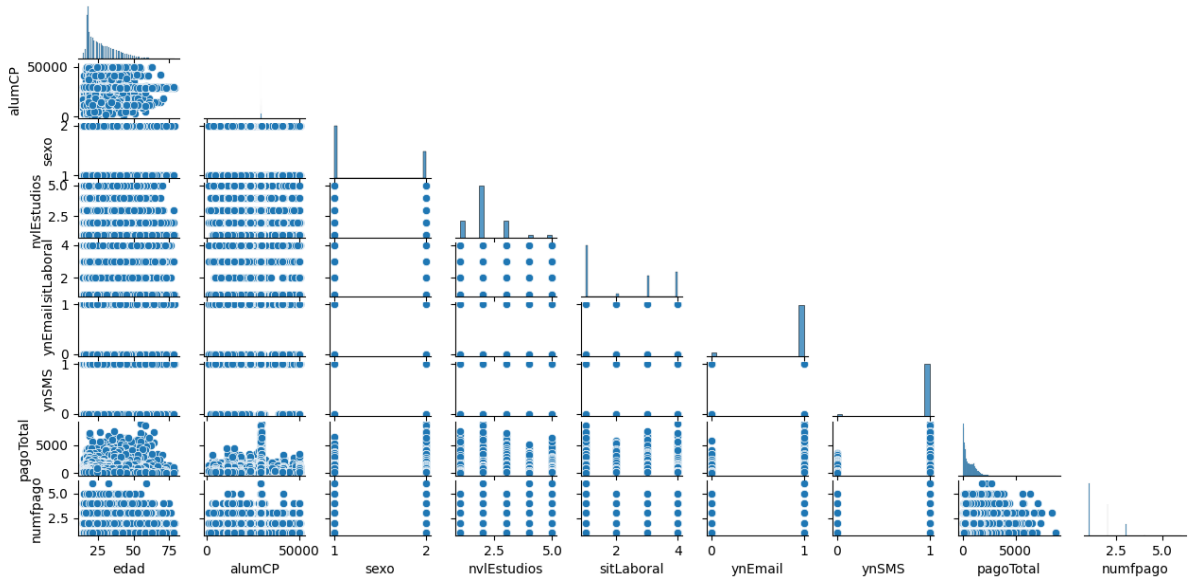

Figura 7.12: Diagrama de correlaciones genérica entre todas las variables

La figura 7.12 muestra un diagrama de correlaciones que está realizado en 2D donde hemos expuesto las variables que eran de mayor interés en contraposición una con otras. Encontramos que cada variable que aparece en ordén en el **eje X** estará en la misma posición en el **eje Y**, siendo respectivamente edad, código postales, sexo, nivel de estudios, situación laboral, la aglomeración de puntos indica un mayor número de clientes identificados con las características de la posición de cada punto.

En la primera correlación que encontramos entre **eje X - edad** y **eje Y - el código postal de los clientes** podemos visualizar cómo la explicación del decrecimiento de los clientes mientras aumentaba su edad se cumple ya que mientras más se acerca el límite de edad máxima el número de apariciones es menor, siendo más abultado y marcado la relación con los códigos postales relacionado con Málaga que forman una especie de línea central superior abultada y con un grupo de edades más variados, las correlaciones con los valores más bajos relacionados con Almería entre otros es bastante menor y con edades más jóvenes. Luego en las relaciones entre estos dos códigos postales hay un gran aumento de las apariciones al igual que al acercarnos a los valores limítrofes superiores relacionado con Zaragoza.

En la segunda correlación **eje X - edad** y **eje Y - sexo**, hay una equidad entre estas dos características pero en valores centrales de edad entre **45-60 años** se produce una disminución de sexo femenino y al aumentar la edad superando los **60 años** vemos un aumento del sexo femenino.

En la tercera correlación siguiendo este orden vertical **eje X - edad** y **eje Y - nivel de estudios** se infiere que hay una equidad general entre las edades y los niveles de estudio que existen pero al aumentar la edad y acercarnos a los límites máximos de edad

el nivel de estudio baja considerablemente al igual que las edades más tempranas hay menos apariciones de los títulos superiores del nivel de estudioo.

En la cuarta correlación **eje X - edad** y **eje Y - situación laboral** se denota una equidad entre las edades y las situaciones laborales posibles.

En la quinta y sexta correlación **eje X - edad** y **eje Y - permisión de emails** o **permisión de sms** se denota una equidad y la correlación es casi la misma.

En la penúltima correlación siguiendo este orden vertical **eje X - edad** y **eje Y - cantidad de transacciones** se ve una progresión positiva con respecto a más edad y más gastos realizados en la empresa excepto en el ultimo tramo de edad que disminuye los gastos bruscamente.

En la última correlación siguiendo este orden vertical **eje X - edad** y **eje Y - distintas formas de pagos utilizadas** se ve una progresión negativa con respecto a más edad y más formas de pago utilizadas habiendo varios valores atípicos del uso de máximo de tipos de formas pagos utilizados en toda la base de datos no teniendo una relación clara con la edad, en el tramo final de edad se usan menos tipos.

En la primera correlación de la segunda columna **eje X - código postales** y **eje Y - sexo** se ve una equidad genérica excepto en los valores que se encuentra cerca del centro de los códigos postales que hay una disminución de las mujeres ocasionado por el aumento de la muestra en esa zona.

En la tercera correlación **eje X - código postales** y **eje Y - nivel de estudios** se infiere que hay una equidad general entre los códigos postales y los niveles de estudio con una diminución de los clientes sin estudios en los relacionados a localizaciones geográficas con códigos postales inferiores cómo Almería entre otras.

En la cuarta correlación **eje X - código postales** y **eje Y - situación laboral** se infiere que hay una equidad general entre los códigos postales y la situación laboral con una disminución clara de los clientes autónomos en los relacionados a localizaciones geográficas con códigos postales inferiores cómo Almería entre otras y en superiores cómo Sevilla entre otras.

En la quinta y sexta correlación **eje X - código postales** y **eje Y - permisión emails** o **permisión de sms** aparece una equidad general entre los códigos postales y estas permisiones excepto con una nimidad de aceptación al envío de sms pasado el centro de los códigos postales.

En la penúltima y última correlación siguiendo este orden vertical **eje X - código postales** y **eje Y - cantidad de transacciones y número de distintas formas de pago** una clara relación entre estos dos diagramas de correlación se forma una imagen muy similar en la que hay un crecimiento inicial en ambos casos luego un pequeño descenso un pico muy marcado cuando llegamos a los valores relacionados con Málaga luego otro descenso y finaliza en un aumento de ambos.

En la tercera columna de correlaciones encontramos una equidad en todos los diagramas de correlación excepto en el penúltimo **eje X - sexo** y **eje Y - transacciones realizadas en euros** donde el sexo femenino aporta más beneficios económicos a la empresa.

En la cuarta columna de correlaciones encontramos una equidad en todos los diagramas de correlación excepto en los dos últimos **eje X - nivel de estudios**, **eje Y - transac-**

ciones realizadas en euros y **número de distintas formas de pago**, en el primer caso se denota que hay una relación inversamente proporcional entre el aumento de nivel de estudios y el gasto realizado en la empresa, exceptuando porque los clientes sin estudios son los segundos que más gastan muy seguidos de los que tienen ESO o equivalente. En el ultimo diagrama no hay un patrón claro entre los estudios y el número distintas de forma de pago utilizados.

En la quinta columna de correlaciones al igual que en las últimas columnas encontramos una equidad en todos los diagramas de correlación excepto en los dos últimos **eje X - situación laboral**, **eje Y - transacciones realizadas en euros** y **número de distintas formas de pago**, en el primer caso vemos cómo los estudiantes son los que realizan el pico de pagos más alto de los clientes y los trabajadores por cuenta ajena los que más pagos realizan, totalmente relacionado con el número de clientes que pertenecen a esta clase. Es digno de estudio cómo los clientes desempleados gastan más que los clientes que son trabajadores autónomos. En cuanto a las distintas formas de pago utilizada los trabajadores por cuenta ajena son los únicos que realizan el máximo luego una igualdad máxima entre los otras clases.

Respecto a las columnas en las que el **eje X** toma los valores de **permisión de sms y de emails**, nos proporciona una visión de que los clientes que permiten el envío de información por estas vías realizan pagos de mayor cantidad a la empresa, pero esta información también depende de que el grupo de clientes que aceptan es alrededor del 90 % de la masa de clientes total.

Los clientes en su gran mayoría tienen una aceptación al envío de información a móvil y correo electrónico, capacitando a la empresa a utilizar esa capacidad, además gastando más en los casos que permiten este envío de información.

En el apartado de edades, se ve una relación clara entre las disminución de clientes en las edades y la aparición en localizaciones geográficas, además de unas zonas muy marcadas por la cantidad de clientes y el mayor rango de edades, el nivel de estudios decae al ser muy jóvenes o muy mayores. Con respecto a los pagos se ve una progresión positiva excepto en el tramo final de más mayores, al contrario sucede en el número de pagos que decrece al aumentar la edad.

En el apartado de los códigos postales se ve una clara predominancia de varias zonas geográficas a la hora de hablar de transacciones y uso de distintos números de pago, unas situaciones laborales excepcionales en unas zonas geográficas y una pequeña variación atípica con respecto al sexo y a la permisión de envío de sms.

En el apartado de niveles de estudios se llega a la conclusión de que a medida de que aumenta el nivel de estudio exceptuando el grupo de clientes sin estudios que es el segundo más concurrido, el gasto en la empresa disminuye.

En el apartado de situación laboral muestran que los estudiantes son los que tienen picos de gastos más altos, siendo los trabajadores por cuenta ajena el grupo de clientes que más transacciones realiza y superando el gasto de los desempleados a los trabajadores autónomos, acercándose a los estudiantes. Además de ser los trabajadores por cuenta ajena los que más tipos distintos de formas de pagos utlizan.

Para finalizar en el apartado sobre los distintos formas de pagos nos encontramos que hay una relación casi proporcional en el aumento de gastos con la disminución de usos de

distintos tipos de formas de pago.

7.4. Agrupaciones de los datos

En esta sección utilizaremos distintos algoritmos de machine learning para agrupar los datos y poder inferir información de las agrupaciones.

Cómo se dividen los grupos puede proporcionar información valiosa para predecir posibles datos en un fúturo, además de una clasificación de los clientes por subgrupos específicos muy beneficiosa para el trato individualizado.

Hemos decidido comenzar por una de las técnicas de agrupación más conocidas, las técnicas jerárquicas, el algoritmo elegido es el aglomerativo, utilizando la distancia euclídea y el método **ward**, este método en vez de calcular la distancia directamente, analiza las varianza de los agrupamientos.

Comenzamos realizando una serie de dendogramas sobre los datos proporcionados para hacer una estimación de los posibles agrupamientos en los que nuestros datos se podrán dividir. En este caso se puede visualizar en los siguientes diagramas cómo los dendograma agrupan los datos en agrupaciones de cinco. La aparición de varios diagramas es debido a que al tener una cantidad de datos extensa este algoritmo y dos variables el análisis se vuelve demasiado complejo y se dilata en el tiempo, por lo que se han analizado datos en grupos de 50.000 datos.

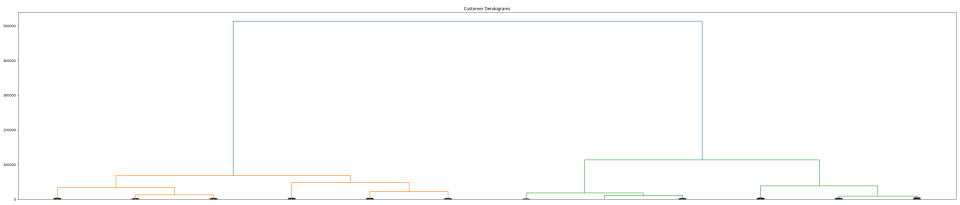

Figura 7.13: Dendograma de las edades de los clientes relacionado con los códigos postales

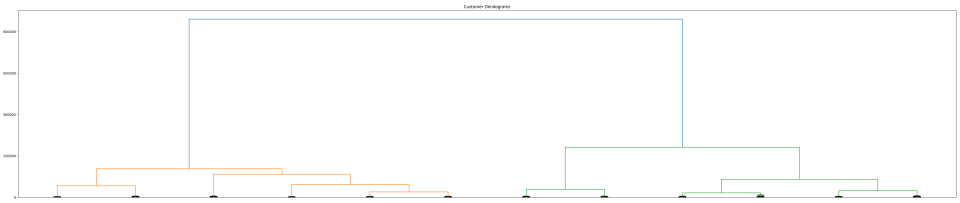

Figura 7.14: Dendograma de las edades de los clientes relacionado con los códigos postales

Después de haber encontrado las estimaciones de los posibles agrupamientos y para validar la mejora sobre otras opciones se hicieron comprobaciones con cuatro agrupamientos, nos decantamos por mostrar las agrupaciones que aparecen al aplicar este algoritmo seleccionando la opción de cinco agrupaciones. En los dendogramas de la figura cómo se comenta en el párrafo anterior se dividen en cuatro y cinco grupos, en la figuras 7.13 , 7.14 y 7.15 se dividen en cinco si se visualiza una línea horizontal sobre las última unión de los árboles formándose 3 grupos en la zona amarilla y 2 en la zona verde.

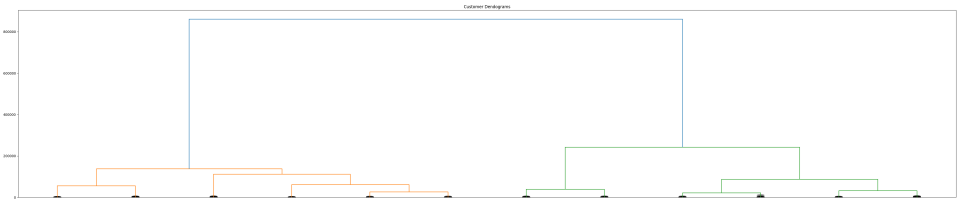

Figura 7.15: Dendograma de las edades de los clientes relacionado con los códigos postales

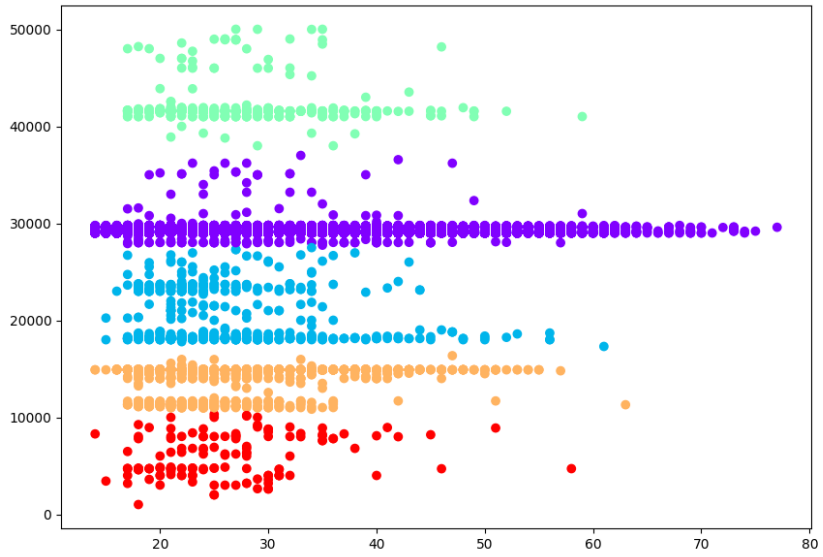

Figura 7.16: Aplicación del algoritmo de agrupación aglomerativo primeros 50.000 valores

Pasamos a analizar la figura 7.16, en el **eje X** encontramos las edades de los clientes y en el **eje Y** encontramos los códigos postales. La acumulación de los puntos indicará el aumento de clientes en esa zona. En estos tres diagramas se muestran los cincos grupos diferenciados desde arriba abajo.

El **grupo 1** primer grupo en el orden vertical, es el grupo más disperso con respecto a las localizaciones geográficas las edades en este grupo aunque son dispersas encontramos una mayoría en el rango entre 18 y 50 años por lo que tiene un rango de edades intermedio.

El **grupo 2** es el grupo más concentrado de todos con respecto a una localización geográfico y con más clientes de todos, este grupo se identifica con la zona de Málaga y alrededores, con respecto a la edad es el que mayor rango de edad abarca desde lo mínimo a lo máximo permitido en las edades.

El **grupo 3** es un grupo algo disperso pero que acumula clientes en cada rango de códigos postales cómo para dar una impresión más compacta que los otros grupos dispersos, es un grupo de clientes más jóvenes se denota una mayor aglomeración de clientes en el rango de edad 18-40 años.

El **grupo 4** es otro grupo similar al segundo muy concentrando con respecto a la localización geográfica pero con un volumen de clientes bastante menor y un rango de

edad un poco menor abarcando también valores de edad mínimos pero llegando hasta los 60 en general y con algún valores aislados en edades superiores.

El **grupo 5** es el último grupo de la subdivisión y se caracteriza por ser el grupo más joven de la subdivisiones siendo un rango disperso al igual que el grupo 1 y el grupo 3 es el menos disperso de todos y su rango de edades más concurrido está entre los 18 y los 30 años.

Añadimos los dos otros gráficos que muestran la confirmación de que las agrupaciones tienen una relación durante toda la distribución de datos al completo.

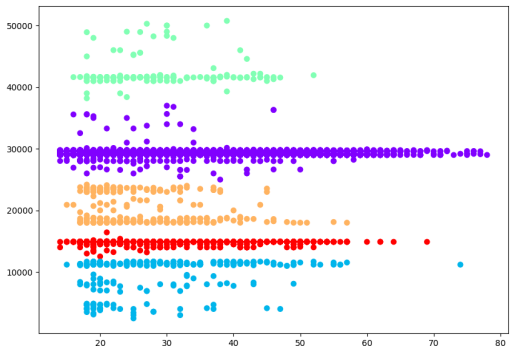

Figura 7.17: Aplicación del algoritmo de agrupación aglomerativo

Figura 7.18: Aplicación del algoritmo de agrupación aglomerativo

Para utilizar distintos tipos de algoritmos de agrupamiento hemos realizado una serie de agrupaciones utilizando el algoritmo de aprendizaje no supervisado K-Means sobre distintos conjunto de datos de este:

La figura 7.19 un gráfico de puntos enfocado a mostrar las secciones grupales de los clientes en base a sus edades, en el **eje X** y el **eje Y** se exponen las edades de los clientes y La unidad de medida son años.

Los puntos rojos que aparecen en la gráfica indican los centroides seleccionado dentro de los datos proporcionados. Posicionados muy cercanos a los siguientes valores: 20,30,40,50 años.

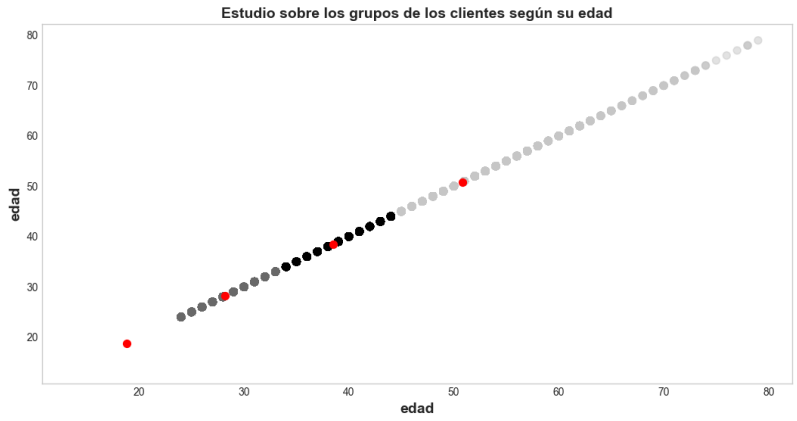

Figura 7.19: Aplicación del algoritmo K-Means sobre las edades

Esto nos deja cuatro agrupaciones claras en base a los datos una agrupación de los más jóvenes con unas edades comprendidas alrededor de los 15 y 25 años. Luego dos agrupaciones de edades intermedias estando entre los 25 y 35 años y los 35 y 45 años y una agrupación mucho más envejecida donde se agrupan el resto de edades ya que los número de clientes que se encuentran en este último rango de edades es mucho menor.

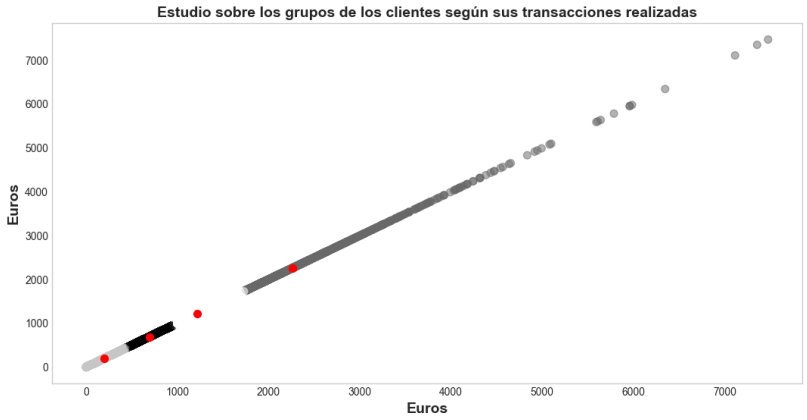

Figura 7.20: Aplicación del algoritmo K-Means sobre los pagos

La figura 7.20 muestra un gráfico de puntos enfocado a mostrar las secciones grupales de los clientes en base a los pagos que han realizado, en el **eje X** y el **eje Y** se exponen las transacciones realizadas por los clientes y La unidad de medida son euros.

Los puntos rojos que aparecen en la gráfica indican los centroides seleccionado dentro de los datos proporcionados. Posicionados muy cercanos a los siguientes valores: 200,700,1200,2300 euros.

Esto nos deja cuatro agrupaciones según rangos económicos progresivos, en el primer grupo se aglomeran una gran parte de las transacciones más frequentes por lo que se crea un grupo con un rango tan delimitado entre 0 y 400. El siguiente grupo que es el segundo más concurrido por los clientes se delimita desde los 400 y los 900 euros. Mientras aumentamos la carga económica el número de clientes que los componen se reducen ese es el motivo por el que estos grupos van aumentando sus rangos delimitado por 900 y

1800 euros y el último grupo con las transacciones más grandes que agrupa todas las superiores a 1800, mucho más disipado en el rango de valores que agrupa pero con un número reducido de clientes.

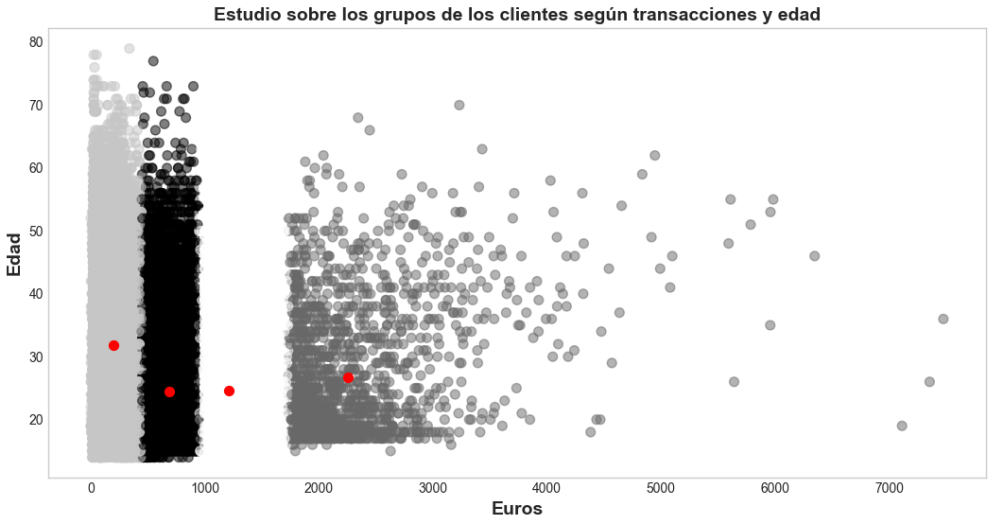

Figura 7.21: Aplicación del algoritmo K-Means sobre edades y pagos

La figura 7.21 muestra un gráfico de puntos enfocado a mostrar las secciones grupales de los clientes en una contraposición de las edades de los clientes y el dinero que han aportado en sus transacciones, en el **eje X** y el **eje Y** se exponen las transacciones realizadas por los clientes y La unidad de medida en la edad son años y en los pagos son euros.

Los puntos rojos que aparecen en la gráfica indican los centroides seleccionado dentro de los datos proporcionados. Posicionados muy cercanos a los siguientes valores: [200 euros,32 años], [700 euros, 24 años], [1200 euros, 25 años] y [2300 euros, 27 años].

Para comenzar se muestra una población de clientes bastante enjuvenecida con respecto al rango de posibles edades que se barajan en las opciones. Encontrándonos con 4 agrupaciones delimitadas claramente.

Una **primera agrupación** más envejecida que el resto en la que encontramos muchos valores de edad más altos que el resto, y un gasto mucho menor que el resto, este grupo es el más poblado de los 4.

En el **segundo grupo** el centroide de edades baja bastante encontrando el grupo con la media más baja, sigue habiendo edades altas pero en casos mucho más aisladas aunque el rango de gastos económicos aumentó.

En la **tercera agrupación** el centroide de las edades casi no varía por lo que la población tiene un rango de edades similares pero en lo que se diferencia es que siendo un grupo bastante joven tienen un rango de gastos económicos más elevado y más ancho.

La **cuarta agruación** destaca por ser un grupo muy disperso aunque sus edades normalmente no sobrepasan a los otros grupos esta dispersión de datos en lo económico produce una mayor concentración de clientes con más edad que en otras situaciones, aunque el centroide de edad no varía mucho de los dos anteriores económicamente hablando recoge el resto de clientes que gastan más de alrededor de 1800 euros.

CAPÍTULO 8

Conclusiones y líneas futuras

8.1. Conclusiones

En este apartado expondremos las deducciones encontradas a partir de todos los estudios, uso de distintas técnicas, herramientas e información de los datos recogidas.

Para comenzar la experiencia con el tratamiendo de datos con una base de datos externa con tantos años de recogida de datos, ha sido bastante complejo debido a que cómo se expone en la sección 4.4, aunque a priori hay una gran cantidad de datos que son de gran valor cualitativo de cara a poder realizar estudios el estado de la base de datos es bastante comprometida con una serie de fallos que ha complicado mucho el tratamiento y el manejo de esta. Además que no traer un manual informativo sobre la misma o la posibilidad de toma de contacto informativo produce una consumición de tiempo en la exploración de la base de datos demasiado grande, focalizándose en los grupos de datos con mayor número de posibles datos que luego mucho de los valores se encontraban corruptos por el formato o por estar vacíos.

En segundo lugar las metodologías desarrolladas en el TFG y usadas como referencia para la ejecución de un proyecto de ciencia de datos ha sido una de los pilares fundamentales, utilizar un método estandarizado y aplicarlo en procesos de ingeniería cómo este es vital para asegurar una ejecución correcta y limpia. Teniendo una gran base informativa por todas partes para apoyarse y recurrir a la hora de trabajar.

En cuanto a las tecnologías y herramientas utilizadas después de comenzar con el uso de R como lenguaje principal, la migración a python ha sido una elección propicia para continuar la formación en lenguajes multiparadigma y mucho más flexible como recurso en otros aspectos, R en cambio tiene limitaciones de uso, aunque ambos tienen un gran potencial para la ciencia de datos python según la experiencia en este TFG tiene más potencial.

A continuación detallaremos las conclusiones con respecto a los resultados hallados al estudiar los conjuntos de datos obtenidos.

En un primer orden de cosas analizaremos los cálculos estadísticos generales donde encontramos en un inicio que predominan los clientes que pagan utilizando un tipo de forma de pago, con un nivel de estudios de ESO o sin estudios, siendo trabajadores por cuenta ajena, hombres y que permiten el envío de publicidad. Al obtener más cálculos

este cuadro se perfila añadiendole más características jóvenes con una edad alrededor de los 27 años, con procedencia Malagueña y un gasto cercano a los 600 euros.

Continuaremos analizando las distribuciones de los datos, el foco principal de clientes son jóvenes de entre 18 y 20 años con un pico muy significativo diferenciándose del resto con un flujo estable aunque menor de clientes alrededor de los 20 y los 40 años. Mostrando un pérfil de clientes afianzado y la posibilidad de atraer a clientes de las franjas menores de edad y mayores de edad con distintos productos.

En el ámbito de localización geográfica la curva se pronuncia claramente para málaga donde se encuentra la mayor parte de los negocios y zona geográfica que lleva más años en activo. Y donde se ve que otras zonas geográficas están empezando a acumular clientes y siendo zonas con potencial de crecimiento.

Después de ir refinando la información nos encontramos con el apartado relacionado con los pagos de los clientes propicios a utilizar pocas distintas formas de pago, siendo la fidelización un objetivo en este aspecto clave. Las transacciones marcan un recorrido muy pronunciado en los pagos comprendidos entre 0 y 200 euros y una estabilidad con bastantes clientes que pagan entre 200 y alrededor de 1000 euros, habiendo una curva negativa en lo que queda hasta el máximo de gastos realizado por los clientes. La compra de varios productos de la empresa debería ser el punto de estudio en estas conclusiones.

En otro orden de cosas analizamos una serie de características cualitativas de los clientes no teniendo estudios o estudios de ESO en su gran mayoría dando la posibilidad a la empresa de generar productos para estos pérfiles específicos o plantear opciones de formación ofertada por esta. Adenmás un gran porcentaje de los clientes son estudiantes o desempleados siendo un pérfil económicamente complejo pero que unos productos enfocados a ellos aumentarían el ingreso de este pérfil de clientes.

La aceptación por la inmensa mayoría de clientes de envíos de emails y sms permite a la empresa plantear campañas de publicidad personalizadas y estratégicas que aumente la fidelización de los clientes y el consumo de distintos productos.

Proseguimos con las correlaciones entre variables, las edades relacionadas a las localizaciones geográficas se trata mejor en las agrupaciones. Los pagos tienen una relación proporcional junto a las edades excepto en el tramo final de las edades, dato atractivo para la captación de clientes con más edad.

A nivel económica las zonas relativas a Málaga son las predominantes aunque varias zonas relativas a otras provincias andaluzas o fuera de la comunidad como Zaragoza empiezan a posicionarse siendo puntos claves para la expansión a nivel económico, sin descuidar la referencia.

Cómo en las conclusiones con respecto a las distribuciones se recoge que el número de clientes disminuía en base a nivel de estudios esto se ve reflejado en el gasto económico que aportan que también disminuye con respecto a su nivel de estudios. En cambio estudiantes y desempleados ocupan un gran volumen de gastos, aunque los trabajadores por cuenta ajena acumulan más transacciones los estudiantes aportan picos de más valor. Los autónomos son un grupo muy reducido. Estas conclusiones pueden ser aprovechadas por la empresa para enfocar productos a estos distintos pérfiles y hacer una comprobación temporal del aumento de ingresos y de aportación económica de estos colectivos.

Por último, el apartado relativo a las agrupaciones de los datos. Comenzamos con la

agrupación de la población respecto a edades y códigos postales. Estas agrupaciones pueden ser utilizadas por la empresa para enfocar estrategías de marketing, productos personalizados, planes de fidelización y de actuación para cada subpoblación.

Subdivide la población en 5 grupos de los cuáles 2 están más concentrados en unas localizaciones geográficas y rangos de edad más extensos por lo que afianzar esas zonas sería una técnica interesante para la empresa. En los otros 3 grupos donde la población está más disipada en cuanto a localización geográfica hay un patrón de edades más jóvenes y más acotada por edades similares, generando la creación de productos personalizados una práctica del interés de la empresa.

Cómo vimos tanto en la representación de distribución como en la agrupación de los datos únicamente por edades, la población se divide en grupo de 10 años, exceptuando los clientes más envejecidos que se podrían agrupar en un solo grupo. Este acotamiento es muy valioso para el enfoque de las campañas publicitarias para captación de clientes según los rangos de edades.

En cuanto a los grupos creados según los rangos económicos del gasto de los clientes predominan los clientes con un gasto reducido, seguidos por gasto medio, gasto grande y gasto muy grande. En este ámbito sería interesante para la empresa analizar la fidelización de los clientes que aportan poco económicamente y los motivos de lo mismo para aumentar el consumo de distintos productos y la confianza de los clientes.

Las últimas agrupaciones con respecto a edades y gasto económico en euros aportan una visión un poco mejor de los tipos de clientes que abarcan la población de la empresa. Una población más envejecida que tiene unos gastos menores y además la población más concurrida, seguida por una población más joven que realiza pagos medios y que está muy concurrida. La siguiente población tiene menor cantidad de individuos pero es joven y tiene unos gastos grandes. Por último el grupo más disperso con respecto a los pagos y a las edades lo que genera una población mediana con respecto a edad, pero que gasta más con respecto a pagos individuales.

Con respecto a los algoritmos aplicados el algoritmo jerárquico aglomerativo tiene mejor eficiencia con valores cuantitativos y una muestra algo menor, en cambio el algoritmo K-Means acepta mejor un mayor número de datos y valores cualitativos.

8.2. Futuras líneas de trabajo

Una de las ventajas de los proyectos de ciencia de datos es su gran polivalencia y la posibilidad de crecer exponencialmente, por lo que habría muchas posibilidades de expandir este trabajo fin de grado.

Para comenzar el siguiente paso en el proyecto sería la aplicación de técnigas de agrupamiento a todas las variables que pueden tener una corelación interesante, estudiando los grupos de población que se crean y concluyendo información atractiva de cara a la empresa.

El siguiente paso sería aplicar técnicas de predicción sobre los datos para poder proporcionar a la empresa información sobre los datos que recogeran en el fúturo antes de que aparezcan, para este caso podemos utilizar distintas formas: Cálcular los ajustes de

las regresiones que aparecen en las distribuciones de correlación y hacer cálculos para el fúturo, aplicar series temporales llegando a analizar los posibles datos que aparecerán en función de semanas, meses o años o aumentar la complejidad utilizando redes neuronales de aprendizaje profundo definido en la parte de fundamentos de la ciencia de datos enfocado a las líneas futuras.

Las opciones en el campo de la ciencia de datos son muy extensas en base a las preferencias de la empresa se puede elegir unas u otras.

Bibliografía

[1] B. Johnson, «IA y más datos de pacientes para arreglar la falta de nuevos fármacos,» *MIT Technology Review*, 2019. dirección: https://www.technologyreview.es/s/11056/ia-y-mas-datos-de-pacientes-para-arreglar-la-falta-de-nuevos-farmacos (visitado 14-11-2020).

[2] E. Alba, C. Blum, P. Asasi y col., *Optimization Techniques for Solving Complex Problems*. WILEY, 2009, ISBN: 9780470411346.

[3] Z. Michalewicz, *Genetic Algorithms + Data Structures = Evolution Programs*, 1st. Springer-Verlag Berlin Heidelberg, 1996, ISBN: 9783540606765.

[4] M. Barker, R. Wilkinson y A. Treloar, «The Australian Research Data Commons,» *Data Science Journal*, n.º 18, pág. 44, 2019. DOI: http://doi.org/10.5334/dsj-2019-044. dirección: https://datascience.codata.org/articles/10.5334/dsj-2019-044/ (visitado 15-11-2020).

[5] D. Stephenson, *Big Data Demystified: How to Use Big Data, Data Science and AI to Make Better Business Decisions and Gain Competitive Advantage*. FT Publishing International, 2018, ISBN: 9781292218113.

[6] J. W. Tukey, «The Future of Data Analysis,» *Annals of Mathematical Statistics*, vol. 33, n.º 1, págs. 1-67, 1962. DOI: 10.1214/aoms/1177704711. dirección: https://projecteuclid.org/euclid.aoms/1177704711 (visitado 16-11-2020).

[7] P. Naur, *Concise Survey of Computer Methods*. Petrocelli Books, 1974, ISBN: 9780884053149.

[8] I. Ladrero. (2020). «10 ejemplos de usos reales de Big Data Analytics,» dirección: https://www.baoss.es/10-ejemplos-usos-reales-big-data/ (visitado 29-11-2020).

[9] J. A. Rodrigo. (2017). «Máquinas de Vector Soporte (Support Vector Machines, SVMs),» dirección: https://www.cienciadedatos.net/documentos/34_maquinas_de_vector_soporte_support_vector_machines (visitado 29-11-2020).

[10] E. S. Brunette, R. C. Flemmer y C. L. Flemmer, «A review of artificial intelligence,» en *2009 4th International Conference on Autonomous Robots and Agents*, localización: Wellington, Nueva Zelanda, Fecha: 10-12 Feb 2009, IEEE, 2009, ISBN: 9781424427123. DOI: 10.1109/ICARA.2000.4804025.

[11] M. Sugiyama, *Introduction to Statistical Machine Learning*. Morgan Kaufmann, 2015, ISBN: 9780128021217.

[12] E. Alba, «Challenges for Real Applications,» Diapositivas de la asignatura Ingeniería y Ciencia de Datos I del Máster U. Ingeniería Informática de la UMA, 2020, dirección: https://www.uma.es/centers/subject/etsi-informatica/5296/102675/ (visitado 16-11-2020).

[13] H. Mason y C. Wiggins. (2010). «A Taxonomy of Data Science.» dataists, ed., dirección: http://www.dataists.com/2010/09/a-taxonomy-of-data-science/ (visitado 15-11-2020).

[14] D. Patil, *Data Jujitsu: The Art of Turning Data into Product*. O'Reilly Media, Inc., 2012, ISBN: 9781449341152.

[15] IBM. (2015). «IBM SPSS Modeler CRISP-DM Guide,» dirección: https://www.ibm.com/support/knowledgecenter/it/SS3RA7_sub/modeler_crispdm_ddita/modeler_crispdm_ddita-gentopic1.html (visitado 16-11-2020).

[16] K. S. Rubin, *Essential Scrum: A Practical Guide to the Most Popular Agile Process—A Practical Guide to the Most Popular Agile Process*. Addison-Wesley Professional, 2012, ISBN: 9780137043293.

[17] D. Allen, *Getting Things Done*. PENGUIN, 2003, ISBN: 9780142000281.

[18] T. I. y Seguridad S.L. (2020). «Web Torcal Formación,» dirección: https://torcal.es/ (visitado 14-11-2020).

UNIVERSIDAD DE MÁLAGA | uma.es

E.T.S. DE INGENIERÍA INFORMÁTICA

E.T.S de Ingeniería Informática
Bulevar Louis Pasteur, 35
Campus de Teatinos
29071 Málaga